# Model-based methods for high-dimensional multivariate analysis

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Aaron J. Molstad

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Adam J. Rothman, Adviser

April 2017

# Acknowledgements

I must first and foremost thank my advisor, Adam J. Rothman. Through our time working together, I have learned a great deal about statistics, the research process, and writing. He has always been thoughtful, patient, and honest in his mentorship. I am incredibly grateful for all Adam has done for my life and career.

I must also thank the faculty and staff in the School of Statistics. In particular, I would like to acknowledge the following faculty and staff members: Xiaotong Shen, Galin Jones, and Dennis Cook for many helpful discussions, for their many letters of recommendation, and for teaching courses which greatly influenced the research in this dissertation; Charlie Geyer for his guidance early in my graduate studies; Julian Wolfson of Biostatistics for our many conversations and his helpful career advice.; Hui Zou for serving on my thesis committee; and Barbara Kuzmak for her tremendous support in my development as a teacher. Thank you to the University of Minnesota for the Doctoral Dissertation Fellowship, which supported the final year of work on this thesis.

Liliana Forzani and Karl Oskar Ekvall deserve acknowledgment for helpful conversations that contributed to third and fourth chapters of this thesis, respectively.

I could not have completed this thesis without the friendship and support of my fellow graduate students. Thank you to Dan Eck, Karl Oskar Ekvall, Adam Maidman, Brad Price, Ben Sherwood, Dootika Vats, and Yang Yang, without whose advice, collaboration, support, and friendship, this would never have been possible.

I am forever grateful to my parents, Brad Molstad and Mary Japs, for their unconditional love and support. Finally, I must also thank Elizabeth Harder for her support and encouragement throughout my graduate studies. I am incredibly lucky to have shared these years with such an inspiring, intelligent, and thoughtful partner.

# Dedication

To my parents.

## Abstract

This thesis consists of three main parts. In the first part, we propose a penalized likelihood method to fit the linear discriminant analysis model when the predictor is matrix valued. We simultaneously estimate the means and the precision matrix, which we assume has a Kronecker product decomposition. Our penalties encourage pairs of response category mean matrix estimators to have equal entries and also encourage zeros in the precision matrix estimator. To compute our estimators, we use a blockwise coordinate descent algorithm. To update the optimization variables corresponding to response category mean matrices, we use an alternating minimization algorithm that takes advantage of the Kronecker structure of the precision matrix. We show that our method can outperform relevant competitors in classification, even when our modeling assumptions are violated. We analyze an EEG dataset to demonstrate our method's interpretability and classification accuracy.

In the second part, we propose a class of estimators of the multivariate response linear regression coefficient matrix that exploits the assumption that the response and predictors have a joint multivariate normal distribution. This allows us to indirectly estimate the regression coefficient matrix through shrinkage estimation of the parameters of the inverse regression, or the conditional distribution of the predictors given the responses. We establish a convergence rate bound for estimators in our class and we study two examples, which respectively assume that the inverse regression's coefficient matrix is sparse and rank deficient. These estimators do not require that the forward regression coefficient matrix is sparse or has small Frobenius norm. Using simulation studies, we show that our estimators outperform competitors.

In the final part of this thesis, we propose a framework to shrink a user-specified characteristic of a precision matrix estimator that is needed to fit a predictive model. Estimators in our framework minimize the Gaussian negative log-likelihood plus an $L_1$ penalty on a linear or affine function evaluated at the optimization variable corresponding to the preci-

sion matrix. We establish convergence rate bounds for these estimators and we propose an alternating direction method of multipliers algorithm for their computation. Our simulation studies show that our estimators can perform better than competitors when they are used to fit predictive models. In particular, we illustrate cases where our precision matrix estimators perform worse at estimating the population precision matrix while performing better at prediction.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Overview

In recent decades, high-dimensional data analysis has emerged as an important area of research in statistics. Enabled by advances in computing and storage, researchers are able to collect more data about each subject than ever before. To make use of this data, statisticians need to develop new statistical methods and algorithms that address the computational and inferential challenges posed by high-dimensionality. They must must also consider whether certain modeling assumptions, such as sparsity, are appropriate or useful for a given application.

In this dissertation, we propose computationally efficient, model-based methods for classification, multivariate response linear regression, and precision (inverse covariance) matrix estimation. The latter two methods acknowledge that in certain applications, the popular assumption of sparsity may not be appropriate or useful and proposes new methods for these cases. Because these methods address different problems, we provide background and a brief review of the relevant literature in each of the subsequent chapters separately.

In Chapter 2, we propose a model-based method for classification when the predictor is matrix-valued, e.g. an image of a handwritten digit. We fit the linear discriminant analysis model by maximizing a penalized matrix-normal log-likelihood. Specifically, we use penalties that encourage zeros in the precision matrix estimator and entrywise equality in pairs of response category mean matrix estimators. For one subproblem of our blockwise-coordinate descent algorithm, we use an alternating minimization algorithm, proposed by Tseng (1991), that scales more efficiently than the majorize-minimize algorithm used to solve

special cases of our subproblem. Our proposed method can be generalized to the quadratic discriminant analysis model and can be adapted for classification when the predictor is a multidimensional array, such as in fMRI or video data. This chapter is based on the work in Molstad and Rothman (2016b).

In Chapter 3, which appears in Molstad and Rothman (2016a), we propose a new method for fitting the multivariate response linear regression model in high-dimensional settings without relying on the popular assumption that the regression coefficient matrix is sparse or has small Frobenius norm. Instead, we assume that predictors and responses have a joint multivariate normal distribution and propose to indirectly estimate the regression coefficient matrix by estimating the conditional distribution of the predictors given the response, i.e, the *inverse regression*. This allows us to fit a parsimonious inverse regression model when the forward regression is not parsimonious. We justify our approach by deriving a convergence rate bound for our indirect estimator. Through extensive simulation studies, we show that our indirect estimator can outperform popular direct estimators in finite samples. Our real data application suggests that our method is especially useful for virtual comparative genomic hybridization (Geng et al., 2011), a method for predicting genetic abnormalities from gene expression data.

In Chapter 4, we propose a new precision matrix estimator for applications when only a characteristic of the precision matrix is needed for prediction. The characteristics we consider are linear or affine functions evaluated at the precision matrix. We propose to estimate the population precision matrix by minimizing the normal negative log-likelihood plus an $L_1$ penalty on the characteristic evaluated at the optimization variable corresponding to the precision matrix. To compute our estimator, we use an alternating direction method of multipliers algorithm that replaces one primal variable update with an approximation based on the majorize-minimize principle. This allows each step of algorithm to be solved in closed form. We also study the statistical properties of our estimator in terms of the sparsity of the characteristic. Specifically, we establish convergence rate bounds for our precision matrix estimator and the characteristic. Unlike existing methods for directly estimating linear or affine functions of the precision matrix, our estimator is applicable to a wide class

of problems in multivariate analysis. In simulation studies, we show that for some linear discriminant analysis data generating models, our method has better classification accuracy than relevant competitors.

# Chapter 2

# A penalized likelihood method for classification with matrix-valued predictors

## 2.1 Introduction

We propose a method for classification when the predictor is matrix valued, e.g. classification of hand-written letters. Standard vector-valued predictor classification methods, such as logistic regression and linear discriminant analysis, could be applied, but they would not take advantage of the matrix structure.

Logistic regression based methods for classification with a matrix-valued predictor have been proposed. Zhou and Li (2014) proposed a nuclear norm penalized likelihood estimator of the regression coefficient matrix $B_* \in \mathbb{R}^{r \times c}$ in a generalized linear model, where the value of the matrix predictor $x \in \mathbb{R}^{r \times c}$ enters the model through the trace of $B_*^{\mathrm{T}} x$. In the same setup, Hung and Wang (2013) assumed that $\mathrm{vec}(B_*) = \beta_* \otimes \alpha_*$ where vec stacks the columns of its argument, $\alpha_* \in \mathbb{R}^r$, $\beta_* \in \mathbb{R}^c$, and $\otimes$ is the Kronecker product. This decomposition was also studied in the dimension reduction literature (Li et al., 2010).

There also exist non-likelihood based methods for classification with a matrix-valued predictor. These approaches modify Fisher's linear discriminant criterion, e.g. 2D-LDA (Li and Yuan, 2005), matrix discriminant analysis (Zhong and Suslick, 2015), and penalized matrix discriminant analysis (Zhong and Suslick, 2015).

We propose a penalized likelihood method for classification with a matrix-valued predictor. Our method estimates the parameters in the linear discriminant analysis model. Let $x_i \in \mathbb{R}^{r\times c}$ be the measured predictor for the $i$th subject and let $y_i \in \{1,\ldots,J\}$ be the measured categorical response for the $i$th subject ($i = 1,\ldots,n$). We assume that $(x_1,y_1),\ldots,(x_n,y_n)$ are a realization of $n$ independent copies of $(X,Y)$ with the following distribution. The marginal distribution of $Y$ is defined by $P(Y=j)=\pi_j$ ($j=1,\ldots,J$), where the $\pi_j$'s are unknown; and

$$\operatorname{vec}(X) \mid Y = j \sim \mathrm{N}_{rc}\left\{\operatorname{vec}\left(\mu_{*j}\right), \Sigma_*\right\}, \quad j = 1,\ldots,J, \tag{2.1}$$

where $\mu_{*j} \in \mathbb{R}^{r\times c}$ is the unknown mean matrix for the $j$th response category, and $\Sigma_*$ is the unknown $rc$ by $rc$ covariance matrix.

We make the simplifying assumption that

$$\Sigma_*^{-1} = \Delta_* \otimes \Phi_*, \tag{2.2}$$

which is equivalent to $\Sigma_* = \Delta_*^{-1} \otimes \Phi_*^{-1}$, where $\Phi_*$ is an unknown $r$ by $r$ precision matrix with $\sum_{a,b}|\Phi_{*a,b}| = r$, and $\Delta_*$ is an unknown $c$ by $c$ precision matrix. The norm condition on $\Phi_*$ is added for identifiability: see Roś et al. (2016) for more on identifiability under (2.2). This simplification of a covariance matrix makes the conditional distributions in (2.1) become matrix normal (Gupta and Nagar, 2000). This exploits the matrix structure of the predictor by reducing the number of parameters in the precision matrix from $O(r^2c^2)$ to $O(r^2+c^2)$.

Several authors have proposed and studied penalized likelihood estimators of $\Phi_*$ and $\Delta_*$ when $J = 1$ (Allen and Tibshirani, 2010; Zhang and Schneider, 2010; Tsiligkaridis et al., 2012; Leng and Tang, 2012; Zhou, 2014).

In this chapter, we propose a penalized likelihood method to fit (2.1) with the assumption in (2.2). Our penalties encourage fitted models that can be easily interpreted by practitioners. We use a blockwise coordinate descent algorithm to compute our estimators. To exploit (2.2) computationally, we use an alternating minimization algorithm

(Tseng, 1991) in one of our block updates. This algorithm scales more efficiently than other popular algorithms, which makes our method computationally feasible for high-dimensional problems. We show that our algorithm has the same computational complexity order as the unpenalized likelihood version, which also requires a blockwise coordinate descent algorithm (Dutilleul, 1999).

## 2.2 Penalized likelihood estimation

### 2.2.1 Proposed method

Let $\mathbb{S}_+^m$ be the set of symmetric and positive definite $m$ by $m$ matrices. The maximum likelihood estimators of the $\mu_{*j}$'s, $\Phi_*$, and $\Delta_*$ minimize the function $g : (\mathbb{R}^{r\times c})^J \times \mathbb{S}_+^r \times \mathbb{S}_+^c \to \mathbb{R}$ defined by

$$g\left(\mu, \Phi, \Delta\right) = \frac{1}{n}\sum_{j=1}^{J}\left[\sum_{i=1}^{n} 1(y_i = j)\mathrm{tr}\left\{\Phi(x_i - \mu_j)\Delta(x_i - \mu_j)^{\mathrm{T}}\right\}\right] - c\log\det(\Phi) - r\log\det(\Delta),$$

where $\mu = (\mu_1, \ldots, \mu_J)$. We propose the penalized likelihood estimators defined by

$$\left(\hat{\mu}, \hat{\Delta}, \hat{\Phi}\right) = \underset{(\mu,\Phi,\Delta)\in\mathcal{T}}{\arg\min}\left\{ g(\mu, \Phi, \Delta) + \lambda_1 \sum_{j<m} \|w_{j,m} \circ (\mu_j - \mu_m)\|_1 + \lambda_2\|\Delta \otimes \Phi\|_1 \right\}, \quad (2.3)$$

$$\text{subject to } \|\Phi\|_1 = r$$

where $\mathcal{T} = (\mathbb{R}^{r\times c})^J \times \mathbb{S}_+^r \times \mathbb{S}_+^c$; $\circ$ is the Hadamard product; $\|\cdot\|_1$ is the sum of the absolute values of the entries of its argument; $\lambda_1$ and $\lambda_2$ are nonnegative tuning parameters; and the $w_{j,m}$'s are $r$ by $c$ user-specified weight matrices.

The first penalty in (2.3) encourages solutions for which pairs of the mean matrix estimates have some equal entries, where this equality occurs in the same locations. Without the first penalty, i.e. $\lambda_1 = 0$, the proposed estimators of the $\mu_{*j}$'s are sample mean matrices. If $\lambda_1 > 0$, then the proposed estimators of the $\mu_{*j}$'s are affected by the estimators of $\Phi_*$ and $\Delta_*$.

We recommend selecting weights similar to those prescribed by Guo (2010). We suggest using $w_{j,m}^{-1} = |\bar{x}_j - \bar{x}_m|, 1 \leq j < m \leq J$ where $\bar{x}_j = \sum_{i=1}^{n} 1(y_i = j)x_i$. Alternatively, one could use weights based on $t$-test statistics or could use weights that incorporate prior information.

The second penalty in (2.3) has a simple impact: for sufficiently large values of $\lambda_2$, some of the entries in the estimate of $\Delta_* \otimes \Phi_*$ are zero, which occurs if and only if either the estimate of $\Delta_*$ or the estimate of $\Phi_*$ has some zero entries. To encourage zeros in estimates of $\Phi_*$ or $\Delta_*$ separately, one could use two separate $L_1$ penalties. Our computational algorithm can be easily adapted to accommodate this case.

The tuning parameters $\lambda_1$ and $\lambda_2$ can be chosen by minimizing the misclassification rate on a validation set.

### 2.2.2 Related work

Xu et al. (2015) proposed fitting the standard linear discriminant analysis model for a vector-valued predictor by penalized likelihood. We can express their parameter estimates in our matrix-predictor setup by setting the number of columns of the matrix predictor to one. Specifically, with $c = 1$ and $\Delta = 1$, Xu et al. (2015) parameter estimates are

$$\underset{(\mu,\Phi)\in(\mathbb{R}^r)^J\times\mathbb{S}_+^r}{\arg\min} \left\{ g(\mu, \Phi, 1) + \lambda_1 \sum_{j<m} \|w_{j,m} \circ (\mu_j - \mu_m)\|_1 + \lambda_2 \sum_{a\neq b} |\Phi_{ab}| \right\}. \tag{2.4}$$

One could view our method as the matrix-valued predictor extension of the method of Xu et al. (2015). Guo (2010) proposed a method that solves a restricted version of (2.4), where $\Phi$ is fixed at a diagonal matrix with pooled sample precision estimates on its diagonal.

Computationally, the algorithms proposed by Xu et al. (2015) and Guo (2010) for solving (2.4) suffer from numerical instability and do not scale efficiently for application to (2.3). In our simulation studies, we compare our proposed method to several competitors, including the method of Guo (2010). The method of Xu et al. (2015) is too slow computationally for the dimensions we consider, so we only use it in a special case when $\Sigma_*$ is known.

## 2.3 Computation

### 2.3.1 Overview

To solve (2.3), we use a block-wise coordinate descent algorithm. Each block update is a convex optimization problem. In the subsequent subsections, we show that updates for $\Phi$ and $\Delta$ can be expressed as the well-studied $L_1$-penalized Gaussian likelihood precision matrix estimation problem. We also use an alternating minimization algorithm for the block update for $\mu$. The algorithm to compute our estimator, along with a set of auxiliary functions, is implemented in the `R` package `MatrixLDA`, which is available on CRAN.

### 2.3.2 Updates for $\Phi$ and $\Delta$

We first derive the update for $\Phi$. Define $\mathrm{GL}(S, \tau)$ as

$$\mathrm{GL}(S, \tau) = \underset{\Theta \in \mathbb{S}_+}{\arg\min} \left\{ \mathrm{tr}(S\Theta) - \log\det(\Theta) + \tau \|\Theta\|_1 \right\}, \tag{2.5}$$

where $S$ is some given nonnegative definite matrix and $\tau$ is a nonnegative tuning parameter. The optimization problem in (2.5) is the $L_1$-penalized Gaussian likelihood precision matrix estimation problem. Many algorithms and efficient software exist to solve (2.5): one good example is the graphical-lasso of Friedman et al. (2008).

Let $f$ be the objective function in (2.3). Suppose $\Delta$ and $\mu$ are fixed. The minimizer of $f$ with respect to $\Phi$ is

$$\tilde{\Phi} = \underset{\Phi \in \mathbb{S}_+^r}{\arg\min} \left\{ \frac{1}{n} \sum_{j=1}^{J} \left[ \sum_{i=1}^{n} 1(y_i = j) \mathrm{tr}\left\{ \Phi(x_i - \mu_j)\Delta(x_i - \mu_j)^{\mathrm{T}} \right\} \right] - c \log\det(\Phi) + \lambda_2 \|\Phi \otimes \Delta\|_1 \right\}. \tag{2.6}$$

Using the fact that $\|\Phi \otimes \Delta\|_1 = \|\Phi\|_1 \|\Delta\|_1$ and

$$\frac{1}{n} \sum_{j=1}^{J} \left[ \sum_{i=1}^{n} 1(y_i = j) \mathrm{tr}\left\{ \Phi(x_i - \mu_j)\Delta(x_i - \mu_j)^{\mathrm{T}} \right\} \right] = c \ \mathrm{tr}\left\{ \Phi S_\phi(\mu, \Delta) \right\},$$

where

$$S_\phi(\mu, \Delta) = \frac{1}{nc} \sum_{j=1}^{J} \left\{ \sum_{i=1}^{n} 1(y_i = j) \, (x_i - \mu_j) \, \Delta \, (x_i - \mu_j)^{\mathrm{T}} \right\},$$

we can express (2.6) as

$$\underset{\Phi \in \mathbb{S}_+^r}{\arg\min} \left[ \operatorname{tr} \{ \Phi S_\phi(\mu, \Delta) \} - \log\det(\Phi) + \frac{\lambda_1 \|\Delta\|_1}{c} \|\Phi\|_1 \right] = \ \mathrm{GL} \left\{ S_\phi(\mu, \Delta), \frac{\lambda_1 \|\Delta\|_1}{c} \right\}.$$

After computing $\tilde{\Phi}$ with $\Delta$ fixed, we can enforce the constraint $\|\Phi\|_1 = r$ using a simple normalization: we replace $(\tilde{\Phi}, \Delta)$ with $(\bar{\Phi}, \bar{\Delta})$, where

$$\bar{\Phi} = \frac{r}{\|\tilde{\Phi}\|_1} \tilde{\Phi}, \quad \bar{\Delta} = \frac{\|\tilde{\Phi}\|_1}{r} \Delta.$$

This ensures that $\|\bar{\Phi}\| = r$ without changing the objective function because $f(\mu, \Delta, \tilde{\Phi}) = f(\mu, \bar{\Delta}, \bar{\Phi})$.

Using a similar argument, the minimizer of $f$ with respect to $\Delta$ with $\mu$ and $\Phi$ fixed is

$$\tilde{\Delta} = \mathrm{GL} \left\{ S_\delta \left( \mu, \Phi \right), \frac{\lambda_1 \|\Phi\|_1}{r} \right\},$$

where

$$S_\delta(\mu, \Phi) = \frac{1}{nr} \sum_{j=1}^{J} \left\{ \sum_{i=1}^{n} 1(y_i = j) \, (x_i - \mu_j)^{\mathrm{T}} \, \Phi \, (x_i - \mu_j) \right\}.$$

### 2.3.3 Update for $\mu$

Let $\Delta$ and $\Phi$ be fixed. The minimizer of $f$ with respect to $\mu$ is

$$\underset{\mu \in \mathbb{R}^{(r \times c) J}}{\arg\min} \frac{1}{n} \sum_{j=1}^{J} \left\{ \sum_{i=1}^{n} 1(y_i = j) \operatorname{tr} \left[ \Phi (x_i - \mu_j) \Delta (x_i - \mu_j)^{\mathrm{T}} \right] \right\} + \lambda_1 \sum_{j<m} \| w_{j,m} \circ (\mu_j - \mu_m) \|_1. \tag{2.7}$$

Special cases of (2.7) have been solved using the majorize-minimize principle (Lange, 2016), where the penalty is majorized by its local-quadratic approximation at the current iterate

(Hunter and Li, 2005). For example, Xu et al. (2015) solved (2.7) when $c = 1$ and $\Delta = 1$; and Guo (2010) solved (2.7) when $c = 1$, $\Delta = 1$, and $\Phi$ was diagonal. However, this majorize-minimize algorithm suffers from numerical instability when iterates for $\mu_j$ and $\mu_m$ are similar from some $(j, m)$. Moreover, if we were to apply the majorize-minimize algorithm to solve (2.7), then each iteration would would have worst case computational complexity $O(r^2c^2)$.

Instead of using an majorize-minimize algorithm, we use an alternating minimization algorithm (Tseng, 1991; Chi and Lange, 2015) to solve (2.7). Our algorithm for solving (2.7) is more numerical stable, each iteration has worst case computational complexity $O(r^2c + c^2r)$, and has a quadratic rate of convergence when implemented with the accelerations proposed by Goldstein et al. (2014). Both the majorize-minimize algorithm and our alternating minimization algorithm require one inversion of $\Delta$ and of $\Phi$.

Similarly to the setup of the alternating direction method of multipliers algorithm (Boyd et al., 2011), we first express (2.7) as a constrained optimization problem:

$$\begin{aligned} &\underset{(\mu,\Theta)\in\mathcal{G}}{\text{minimize}} \; g(\mu, \Phi, \Delta) + \lambda_1 \sum_{j<m} \|w_{j,m} \circ \Theta_{j,m}\|_1 \\ &\quad \text{subject to } \Theta_{j,m} = \mu_j - \mu_m \quad 1 \le j < m \le J, \end{aligned} \tag{2.8}$$

where $\mathcal{G} = \mathbb{R}^{(r\times c)J} \times \mathbb{R}^{(r\times c)J(J-1)/2}$ and $\Theta = (\Theta_{1,2}, \ldots, \Theta_{J-1,J})$ . The augmented Lagrangian for (2.8), using notation similar to Chi and Lange (2015), is

$$\begin{aligned} \mathcal{F}_\rho(\mu, \Theta, \Gamma) =& g(\mu, \Phi, \Delta) + \lambda_1 \sum_{j<m} \|w_{j,m} \circ \Theta_{j,m}\|_1 \\ &+ \sum_{j<m} \operatorname{tr}\left\{\Gamma_{j,m}^{\mathrm{T}} (\Theta_{j,m} - \mu_j + \mu_m)\right\} + \frac{\rho}{2} \sum_{j<m} \|\Theta_{j,m} - \mu_j + \mu_m\|_F^2, \end{aligned}$$

for step size parameter $\rho > 0$ and Lagrangian variables $\Gamma_{j,m} \in \mathbb{R}^{r\times c}$ for $1 \le j < m \le J$. Letting the superscript $t$ denote the value of the $t$-th iterate of an optimization variable,

the alternating minimization algorithm updating equations are

$$\mu^{(t+1)} = \underset{\mu \in \mathbb{R}^{(r\times c)J}}{\arg\min}\, \mathcal{F}_0\left(\mu, \Theta^{(t)}, \Gamma^{(t)}\right), \tag{2.9}$$

$$\Theta^{(t+1)} = \underset{\Theta \in \mathbb{R}^{(r\times c)J(J-1)/2}}{\arg\min}\, \mathcal{F}_\rho\left(\mu^{(t+1)}, \Theta, \Gamma^{(t)}\right), \tag{2.10}$$

$$\Gamma_{j,m}^{(t+1)} = \Gamma_{j,m}^{(t)} + \rho\left(\Theta_{j,m}^{(t+1)} - \mu_j^{(t+1)} + \mu_m^{(t+1)}\right) \text{ for } 1 \le j < m \le J,$$

until convergence. The alternating direction method of multipliers algorithm modifies (2.9) by using $\mathcal{F}_\rho$ rather than $\mathcal{F}_0$. The advantage of using $\mathcal{F}_0$ is that we avoid solving an $rc \times rc$ linear system of equations at complexity $O(r^2c^2)$ when using the Kronecker structure. Using $\mathcal{F}_0$ also allows the updates for $\mu_1, \ldots, \mu_J$ to be computed in parallel with closed form solutions for each. Two conditions for the convergence of alternating minimization are that $g$ is strongly convex (Tseng, 1991), which it is in our case, and that $\rho$ is sufficiently close to zero. We provide a computable bound on the step size $\rho$ to ensure convergence of our alternating minimization algorithm in the subsequent section.

The computational advantage of alternating minimization over alternating direction method of multipliers was also recognized by Chi and Lange (2015) in the context of convex clustering. They found that the simplification of (2.9) relative to the alternating direction method of multipliers version yielded a substantially more efficient algorithm.

Using the first order optimality condition for (2.9),

$$\mu_j^{(t+1)} = \bar{x}_j + \frac{1}{2\hat{\pi}_j}\Phi^{-1}\left(\sum_{\{m:m>j\}} \Gamma_{j,m}^{(t)} - \sum_{\{m:m<j\}} \Gamma_{m,j}^{(t)}\right)\Delta^{-1} \quad j = 1, \ldots, J, \tag{2.11}$$

where $\hat{\pi}_j = n_j/n$ for $j = 1, \ldots, J$.

The zero subgradient equation for (2.10) is

$$\rho\Theta_{j,m}^{(t+1)} + \Gamma_{j,m}^{(t)} - \rho\left(\mu_j^{(t+1)} - \mu_m^{(t+1)}\right) + \left\{\lambda_1 w_{j,m} \circ h\left(\Theta_{j,m}^{(t+1)}\right)\right\} = 0, \tag{2.12}$$

where $h : \mathbb{R}^{r\times c} \to \mathbb{R}^{r\times c}$ and for all $(s,t) \in \{1,\dots r\} \times \{1,\dots,c\}$,

$$[h(x)]_{s,t} = \begin{cases} \text{sign}(x_{s,t}) & : x_{s,t} \neq 0 \\ [-1,1] & : x_{s,t} = 0 \end{cases}.$$

Tibshirani (1996), among others, have shown that (2.12) can be solved using the soft-thresholding operator: $\text{soft}(x,\tau) = \max(|x| - \tau, 0)\text{sign}(x)$. The update for $\Theta_{j,m}$ is

$$\Theta_{j,m}^{(t+1)} = \text{soft}\left(\mu_j^{(t+1)} - \mu_m^{(t+1)} - \rho^{-1}\Gamma_{j,m}^{(t)}, \frac{\lambda_1}{\rho} w_{j,m}\right),$$

where soft is applied elementwise.

We use an accelerated variation of the algorithm presented in this section to solve (2.7). This is based on Goldstein et al. (2014) with simple restarting rules described by O'Donoghue and Candes (2015). Further details about our implementation are given in the subsequent section.

### 2.3.4 Summary

The block-wise coordinate descent algorithm for solving (2.3) is summarized in Algorithm 1.

**Algorithm 1** Blockwise coordinate descent algorithm for (2.3)

Initialize $\Delta^{(0)} \in \mathbb{S}_c^+$, $\Phi^{(0)} \in \mathbb{S}_r^+$ such that $\|\Phi^{(0)}\|_1 = r$. Set $m = 0$. Repeat Step 1 - 5 until convergence.

*Step 1.* Compute $\mu^{(m+1)} = \arg\min_{\mu\in\mathbb{R}^{(r\times c)J}} g\left(\mu, \Phi^{(m)}, \Delta^{(m)}\right) + \lambda_1 \sum_{j<m} \|w_{j,m} \circ (\mu_j - \mu_m)\|_1$ using the algorithm in Section 2.3.3;
*Step 2.* Compute $\tilde{\Delta} = \text{GL}\left\{S_\delta\left(\mu^{(m+1)}, \Phi^{(m)}\right), \lambda_2\right\}$;
*Step 3.* Compute $\tilde{\Phi} = \text{GL}\left\{S_\phi\left(\mu^{(m+1)}, \tilde{\Delta}\right), \frac{\lambda_2}{c}\|\tilde{\Delta}\|_1\right\}$;
*Step 4.* Set $\Delta^{(m+1)} = \frac{\|\tilde{\Phi}\|_1}{r}\tilde{\Delta}$, $\Phi^{(m+1)} = \frac{r}{\|\tilde{\Phi}\|_1}\tilde{\Phi}$;
*Step 5.* Replace $m$ with $m+1$.

To get initial values $\Phi^{(0)}$ and $\Delta^{(0)}$, we run the maximum likelihood algorithm (Dutilleul, 1999) until a mild convergence tolerance is reached, and use $\Phi^{(0)} = \text{diag}(\Phi^{\text{MLE}})$ and $\Delta^{(0)} = \text{diag}(\Delta^{\text{MLE}})$ where $\left(\Phi^{\text{MLE}}, \Delta^{\text{MLE}}\right)$ are the final iterates.

Let $k_\phi^{(m)} = \varphi_{\min}(\Phi^{(m)})$ and $k_\delta^{(m)} = \varphi_{\min}(\Delta^{(m)})$, where $\varphi_{\min}(\cdot)$ denotes the minimum eigenvalue of its argument. For the $(m+1)$th update of $\mu$, if we select the step size parameter

$$\rho^{(m+1)} \in \left(0, \left\{\min_j \{\hat{\pi}_j\} 4k_\phi^{(m)} k_\delta^{(m)}\right\} / J\right), \tag{2.13}$$

then the alternating minimization algorithm converges (Tseng, 1991; Chi and Lange, 2015). One can verify that (2.7) and (2.13) satisfy the conditions for convergence stated in Section 6.2 of the Supplemental Material of Chi and Lange (2015) using an argument similar to theirs. The minimum eigenvalues of $\Phi^{(m)}$ and $\Delta^{(m)}$ are positive as long as initializers $\Phi^{(0)}$ and $\Delta^{(0)}$ are positive definite. When $k_\delta^{(m)}$ and $k_\phi^{(m)}$ are positive, $g$ is strongly convex in $\mu$, which is required for convergence.

In practice, we find it better to use $\rho$ an order of magnitude smaller than the upper bound in (2.13), i.e., we use $\rho^{(m+1)} = (\min_j \{\hat{\pi}_j\} 4k_\phi^{(m)} k_\delta^{(m)})/(10J)$ to ensure numerical stability. Although the step size $\rho^{(m+1)}$ may be small when $\Phi^{(m)}$ and $\Delta^{(m)}$ are dense, we find that when using accelerations and warm-starts, the small step size is not problematic.

We use an accelerated version of the alternating minimization algorithm proposed by Goldstein et al. (2014), which was also used by Chi and Lange (2015). O'Donoghue and Candes (2015) showed that acceleration restarts imposed after a fixed number of iterations can decrease the number of iterations required for convergence. In our implementation of the alternating minimization algorithm, we restart the accelerations after 200 iterations. We warm-start the $(m+1)$th update of $\mu$ by initializing the Lagrangian variables at their final iterates from the $m$th update.

At convergence of the alternating minimization algorithm, zeros in the final iterate of $\Theta_{j,m}$ do not correspond to exact entrywise equality in the final iterates for $\mu_j$ and $\mu_m$. To enforce equality at the solution, we use simple thresholding.

### 2.3.5 Computational complexity

Solving (2.3) with $\lambda_1 = \lambda_2 = 0$, i.e. maximum likelihood estimation, also requires a blockwise coordinate descent algorithm (Dutilleul, 1999). The maximum-likelihood blockwise coordinate descent algorithm has computational complexity of order $O(nr^2c + nc^2r + r^3 + c^3)$. The first two terms come from computing the sample covariance matrices $S_\phi$ and $S_\delta$, and the last two terms come from inverting $S_\phi$ and $S_\delta$.

Our algorithm's computational complexity is also $O(nr^2c + nc^2r + r^3 + c^3)$. We compute $S_\phi$ and $S_\delta$ and the graphical-lasso algorithm that we use is known to have worst case complexity $O(p^3)$ for a estimating a $p \times p$ precision matrix (Witten et al., 2011). In addition, for each $\mu$ update, we compute eigendecompositions of the iterates for $\Phi$ and $\Delta$. The alternating minimization algorithm costs $O(r^2c + c^2r)$ when implemented in parallel.

The magnitude of tuning parameters effects the computing time of our algorithm. Generally, smaller values of $\lambda_2$ take longer.

## 2.4 Simulations

### 2.4.1 Models

For 100 independent replications, we generated a realization of $n = n_{\text{train}} + n_{\text{validate}} + n_{\text{test}}$ independent copies of $(X, Y)$, where we set $n_{\text{train}} = n_{\text{validate}} = 75$, and $n_{\text{test}} = 1000$. The categorical response $Y$ has support $\{1, 2, 3\}$ with probabilities $\pi_{*1} = \pi_{*2} = \pi_{*3} = 1/3$. Then

$$\text{vec}\,(X) \mid Y = j \sim \text{N}_{rc}\left\{\text{vec}\,(\mu_{*j}), \Sigma_*\right\},$$

where $\mu_{*1}, \mu_{*2}$, and $\mu_{*3}$ are only different in one $4 \times 4$ submatrix, whose position is chosen randomly in each replication. We used multiple choices for the entries in this submatrix, which are displayed in Figure 2.1. All other mean matrix entries were set to zero. We consider four covariance models:

*Model 1.* $\Sigma_* = \Delta_* \otimes \Phi_*$ where $\Phi_*$ has $(a, b)$th entry $0.7^{|a-b|}$ and $\Delta_*$ has $(c, d)$th entry $0.7 \times 1(c \neq d) + 1(c = d)$.

*Model 2.* $\Sigma_* = \Delta_* \otimes \Phi_*$ where $\Phi_*$ has $(a, b)$th entry $0.7^{|a-b|}$ and $\Delta_*$ is block-diagonal where $\Delta_*$ can be expressed elementwise:

$$\Delta_{c,d} = \begin{cases} 1 & \text{if } c = d \\ 0.7 & \text{if } \mu_{*j,a,c} \neq \mu_{*m,a,d} \text{ for any } a \in \{1, \ldots, r\} \text{ and } 1 \leq j < m \leq J \\ 0 & \text{otherwise} \end{cases}.$$

*Model 3.* $\Sigma_*$ corresponds to the covariance model

$$\text{Cov}(X_{a,b}, X_{c,d} \mid Y = j) = \{0.5I(b \neq d) + I(b = d)\} \frac{(\rho_b \rho_d)^{|a-c|}}{1 - \rho_b \rho_d},$$

where $\rho_1, \ldots, \rho_c$ are $c$ equally spaced values between 0.5 and 0.9. The matrix $\Sigma_*$ is positive definite when $r = c$ with $c = \{8, 16, 32, 64\}$, and when $r = 32$ with $c = \{8, 16, 32, 64\}$.

*Model 4.* $\Sigma_*$ corresponds to the covariance model

$$\text{Cov}(X_{a,b}, X_{c,d} \mid Y = j) = \begin{cases} 1 & \text{if } (a, b) = (c, d) \\ 0.5 & \text{if } \mu_{*j,a,b} \neq \mu_{*m,c,d} \text{ for any } 1 \leq j < m \leq J \\ 0 & \text{otherwise} \end{cases}.$$

In Model 3, if $\rho_k = \rho$ for all $k \in \{1, \ldots, c\}$, then $\Sigma_*$ has the decomposition (2.2) corresponding to $\Phi_*$ with an AR(1) structure and $\Delta_*$ with a compound symmetric structure (Mitchell et al., 2006). However, when $\rho_k \neq \rho$, $\Sigma_*$ does not have decomposition (2.2): the correlation between any two entries in the same row depends on the column and vice versa. Model 4 is the $rc-$variate normal model similar to the first model used in the simulations from Xu et al. (2015).

### 2.4.2 Methods

We consider the following model-based methods for fitting the linear discriminant analysis model:

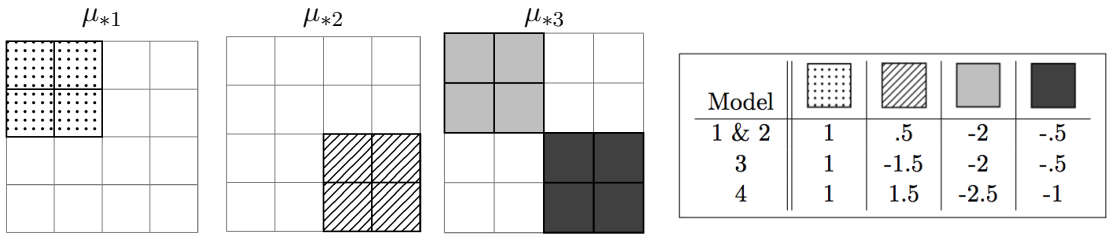



Figure 2.1: The $4 \times 4$ submatrix where $\mu_{*1}$, $\mu_{*2}$, and $\mu_{*3}$ differ. White corresponds to zero and the legend gives the values corresponding to the highlighted cells for each model.

- Bayes. The Bayes rule, i.e., $\Sigma_*$, $\mu_*$, and $\pi_{*j}$ known for $j = 1, \ldots, J$;
- MN. The maximum likelihood estimator of (2.1) under (2.2), i.e., the matrix-normal maximum likelihood estimator;
- Guo. The sparse naïve Bayes type-estimator proposed by Guo (2010) with tuning parameter chosen to minimize misclassification rate on the validation set;
- vec-SURE. The SURE independence screening method proposed by Pan et al. (2016) with model sizes chosen to minimize misclassification rate on the validation set;
- MN-SURE. The matrix-normal extension of the SURE independence screening estimator proposed by Pan et al. (2016) with model sizes chosen to minimize misclassification error on the validation set.
- PMN($\mu$). The estimator defined by (2.3) with $\mu = \mu_*$ fixed and $\lambda_2$ chosen to minimize misclassification rate on the validation set;
- PMN($\Sigma$) / Xu($\Sigma$). The estimator defined by (2.3) with $\Phi = \Phi_*$ and $\Delta = \Delta_*$ fixed when $\Sigma_* = \Delta_* \otimes \Phi_*$; the estimator defined by (2.4) with $\hat{\Sigma} = \Sigma_*$ fixed when $\Sigma_* \neq \Delta_* \otimes \Phi_*$; and $\lambda_1$ chosen to minimize misclassification rate on the validation set;
- PMN. The estimator defined by (2.3) with tuning parameters chosen by minimizing misclassification rate on the validation set.

The methods PMN($\mu$) and PMN($\Sigma$) / Xu($\Sigma$) both use some oracle information and were included to study how estimating $\mu_*$, $\Delta_*$, and $\Phi_*$ simultaneously affect classification accu-

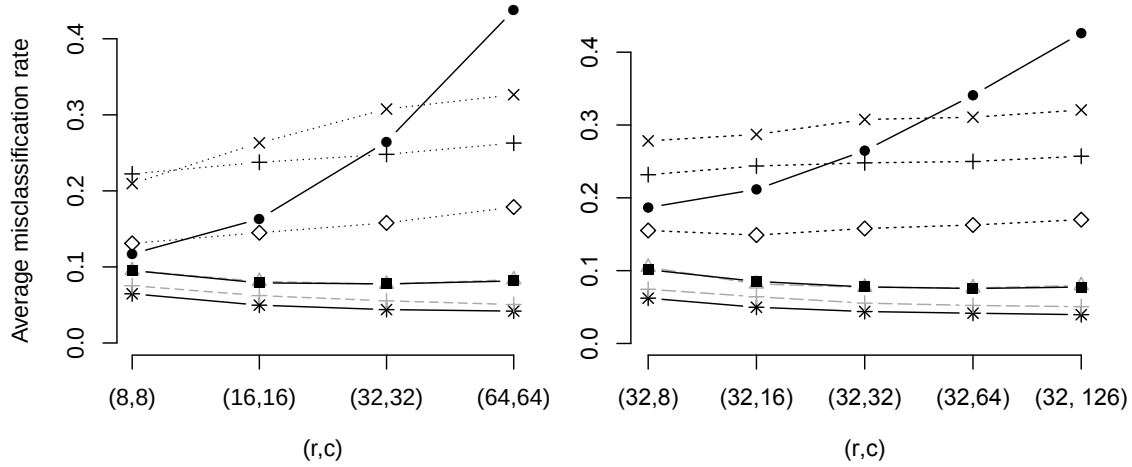
racy. We refer to these method as part-oracle matrix-LDA methods. We refer to Guo and vec-SURE as vector-LDA methods; MN and MN-SURE as non-oracle matrix-LDA methods. MN-SURE is a matrix-normal generalization of the screening method proposed by Pan et al. (2016).

Following Guo (2010), we use a validation set to select tuning parameters. The candidate set for tuning parameters was $\{2^x : x = -12, -11.5, \ldots, 11.5, 12\}$. Candidate model sizes for vec-SURE and MN-SURE were $\{0, 1, \ldots, 25\}$, where model size refers to the number of pairwise nonzero mean differences based on thresholding.

Table 2.1: True negative rates and true positive rates, respectively, averaged over the 100 replications for the models from Section 2.4.1.

| | Model 1 $(r, c)$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Method | (8,8) | (16,16) | (32,32) | (64,64) | (32,8) | (32,16) | (32,64) | (32,126) |
| Guo | 85.7/79.4 | 95.8/68.8 | 98.6/65.6 | 99.4/59.8 | 96.8/70.9 | 97.6/68.2 | 99.2/65.5 | 99.5/59.4 |
| vec-SURE | 88.5/71.9 | 97.9/52.4 | 99.7/40.0 | 99.8/35.4 | 98.4/49.9 | 99.2/48.1 | 99.8/38.9 | 99.9/35.2 |
| MN-SURE | 35.2/90.9 | 80.2/66.9 | 97.3/46.5 | 99.2/37.9 | 87.1/64.8 | 90.0/61.5 | 98.4/43.6 | 99.4/38.9 |
| PMN($\Sigma$) | 85.9/88.2 | 94.0/84.1 | 98.2/78.4 | 99.0/81.2 | 94.1/82.6 | 96.7/83.5 | 98.7/80.6 | 99.2/74.9 |
| PMN | 95.1/79.9 | 95.8/77.5 | 99.0/74.0 | 99.5/69.9 | 98.2/74.2 | 98.6/74.6 | 99.3/73.9 | 99.6/71.6 |
| | Model 2 $(r, c)$ | | | | | | | |
| | (8,8) | (16,16) | (32,32) | (64,64) | (32,8) | (32,16) | (32,64) | (32,126) |
| Guo | 81.4/81.6 | 94.7/73.2 | 97.5/65.8 | 98.6/60.8 | 94.6/74.1 | 96.6/68.8 | 99.0/62.5 | 99.0/63.1 |
| vec-SURE | 87.1/71.2 | 98.1/51.1 | 99.7/36.2 | 99.9/29.5 | 98.1/51.4 | 99.2/47.0 | 99.8/35.5 | 99.9/32.2 |
| MN-SURE | 46.1/89.0 | 74.2/72.9 | 91.8/54.0 | 97.6/42.9 | 83.9/68.9 | 86.4/65.5 | 96.9/43.9 | 98.2/39.4 |
| PMN($\Sigma$) | 90.9/87.5 | 94.4/86.9 | 98.8/80.6 | 99.6/84.2 | 95.0/86.9 | 97.4/85.5 | 99.2/82.5 | 99.4/79.9 |
| PMN | 96.5/79.1 | 96.9/77.5 | 99.1/73.0 | 99.8/70.5 | 98.7/77.6 | 99.1/74.2 | 99.5/68.9 | 99.7/70.9 |
| | Model 3 $(r, c)$ | | | | | | | |
| | (8,8) | (16,16) | (32,32) | (64,64) | (32,8) | (32,16) | (32,64) | (32,126) |
| Guo | 86.8/84.2 | 95.5/81.2 | 97.9/75.1 | 99.4/62.6 | 95.3/80.4 | 96.0/79.8 | 98.7/70.2 | —/— |
| vec-SURE | 84.8/82.9 | 97.2/65.5 | 99.4/44.6 | 99.8/31.2 | 97.8/55.5 | 98.8/52.6 | 99.7/37.0 | —/— |
| MN-SURE | 38.6/94.2 | 80.6/81.5 | 96.4/53.2 | 98.8/35.5 | 85.5/72.8 | 90.8/67.4 | 97.7/43.0 | —/— |
| Xu($\Sigma$) | 81.4/92.1 | 93.8/90.0 | 96.3/86.9 | 98.5/83.9 | 91.4/87.1 | 95.7/87.9 | 97.9/87.1 | —/— |
| PMN | 86.4/96.1 | 93.8/96.0 | 98.3/93.0 | 99.3/87.2 | 93.8/93.5 | 95.7/94.1 | 98.8/90.1 | —/— |
| | Model 4 $(r, c)$ | | | | | | | |
| | (8,8) | (16,16) | (32,32) | (64,64) | (32,8) | (32,16) | (32,64) | (32,126) |
| Guo | 82.2/98.5 | 93.9/98.5 | 96.9/97.1 | 98.9/96.1 | 90.4/98.2 | 95.0/97.9 | 97.8/93.4 | 98.9/96.0 |
| vec-SURE | 97.7/83.1 | 99.4/79.9 | 99.8/78.4 | 99.9/70.6 | 99.2/80.1 | 99.7/79.1 | 99.9/74.6 | 99.9/74.9 |
| MN-SURE | 73.1/97.1 | 89.8/96.2 | 96.0/91.0 | 98.8/82.4 | 89.9/94.8 | 95.7/91.6 | 98.4/84.5 | 99.0/84.4 |
| Xu($\Sigma$) | 87.7/99.2 | 93.3/98.4 | 97.8/96.9 | 99.5/97.1 | 92.0/97.8 | 96.3/96.2 | 98.8/95.8 | 99.4/96.1 |
| PMN | 92.6/96.6 | 95.8/97.5 | 97.3/94.9 | 99.5/92.2 | 94.3/96.4 | 97.6/95.5 | 99.2/91.5 | 99.4/93.8 |

(a) Model 1 with $r = c$ (b) Model 1 with $r = 32$




(c) Model 2 with $r = c$ (d) Model 2 with $r = 32$

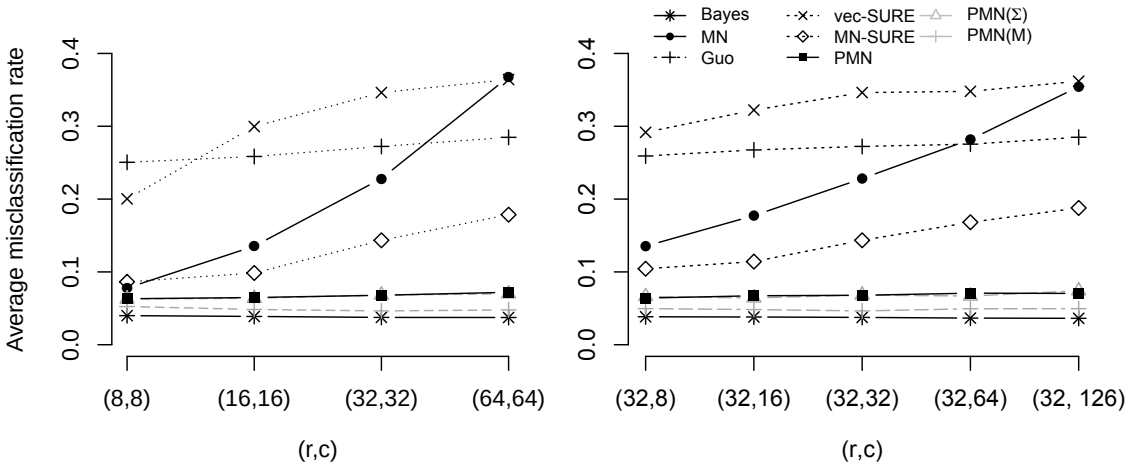



Figure 2.2: Misclassification rates averaged over 100 replications; (a) and (b) are for Model 1 and (c) and (d) for Model 2.

### 2.4.3 Performance measures

To compare classification accuracy, we record the misclassification rate on the test set for each replication. We also measure identification of mean differences that are zero through both true positive and true negative rate. Let $D(\mu_*) = \left[\text{vec}(\mu_{*1} - \mu_{*2}), \ldots, \text{vec}(\mu_{*(J-1)} - \mu_{*J})\right]$, and $D(\hat{\mu}) = \left[\text{vec}(\hat{\mu}_1 - \hat{\mu}_2), \ldots, \text{vec}(\hat{\mu}_{(J-1)} - \hat{\mu}_J)\right]$. We define the true positive rate of an estimator $\hat{\mu}$ as

$$\frac{\text{card}\left\{(z,w) : [D(\hat{\mu})]_{z,w} \neq 0 \cap [D(\mu_*)]_{z,w} \neq 0\right\}}{\text{card}\left\{(z,w) : [D(\mu_*)]_{z,w} \neq 0\right\}},$$

where card denotes cardinality of a set. We similarly define the true negative rate of an estimator $\hat{\mu}$ as

$$\frac{\text{card}\left\{(z,w) : [D(\hat{\mu})]_{z,w} = 0 \cap [D(\mu_*)]_{z,w} = 0\right\}}{\text{card}\left\{(z,w) : [D(\mu_*)]_{z,w} = 0\right\}}.$$

True positive and true negative rates together address mean difference estimation which we use as a measure of variable selection for comparison to the estimator of Guo (2010) and Pan et al. (2016).

### 2.4.4 Results

We display average misclassification rates for Models 1 and 2 in Figure 2.2. For Model 1, the matrix-normal maximum likelihood estimator tended to outperform the vector-LDA methods when $r$ and $c$ were small, but its average classification rate got worse as the dimensionality increases. The estimator proposed by Guo (2010) performs poorly when $r$ and $c$ are small, but got worse more slowly than the other vector and non-oracle matrix-LDA methods. The misclassification rate of the Bayes rules suggests that as the dimensionality increases in Model 1, the optimal misclassification rate can be improved. Our method PMN had improved classification accuracy as both $r$ and $c$ increased and performed similarly to PMN($\Sigma$), which uses some oracle information.

True positive and true negative rate results are displayed in Table 2.1. For Model 1, PMN tended to have the second highest true negative rate behind vec-SURE, but tends

to have higher true positive rate than all competing methods except PMN($\Sigma$), which uses some oracle information.

Results were similar for Model 2. The matrix-normal variation of the SURE screening estimator of Pan et al. (2016) tended to perform best among the vector and non-oracle matrix-LDA methods. The estimator of Guo (2010) got worse the slowest amongst the vector-LDA methods. PMN performed as well as PMN($\Sigma$), both of which performed more closely to PMN($\mu$) and the Bayes rule than for Model 1.

The misclassification rates for Models 3 and 4 are displayed in Figure 2.3. In Model 3, although $\Sigma_*$ does not have the Kronecker decomposition in (2.2), PMN outperformed all non part-oracle estimators. In terms of true positive and true negative rate results presented in Table 2.1, PMN performed similarly to Xu($\Sigma$), both of which had higher true positive rate than competitors and true negative rate similar to vec-SURE. This suggests that even when (2.2) does not hold, our method can perform well in classification.

In Model 4, PMN performed similarly to the vector-LDA methods. MN-SURE was the best non-oracle method, which suggests that (2.2) may be a reasonable alternative to naïve Bayes under high dimensionality. Like in Model 3, PMN performed similarly to Xu($\Sigma$) in terms of true positive and true negative rates.

We also report timing results for all the settings in Model 1. In our simulations, we used a $12 \times 12$ grid of candidate tuning parameters $(\lambda_1, \lambda_2)$ for our method. For each point on the grid, we compute the average computing time over the 100 replications. In Table 2.2 we report the minimum, maximum, mean, and median average computing times for all of the points on the grid computing using warm-starts. We also include median and mean average computing times (without warm-starts) for the tuning parameter pair selected by minimizing the misclassification rate on the validation set. In Figure 2.4, we show smoothed contour plots of average computing times for the cases where $r = 32$ and $c = 32$ and $c = 64$.

All computations were performed on an Intel Core i7-3770 CPU at 3.4GHz with 8GB of RAM. Our package `MatrixLDA` was designed for computation on a single CPU, but the source code can be easily modified to allow for parallelization, which could reduce computing times.

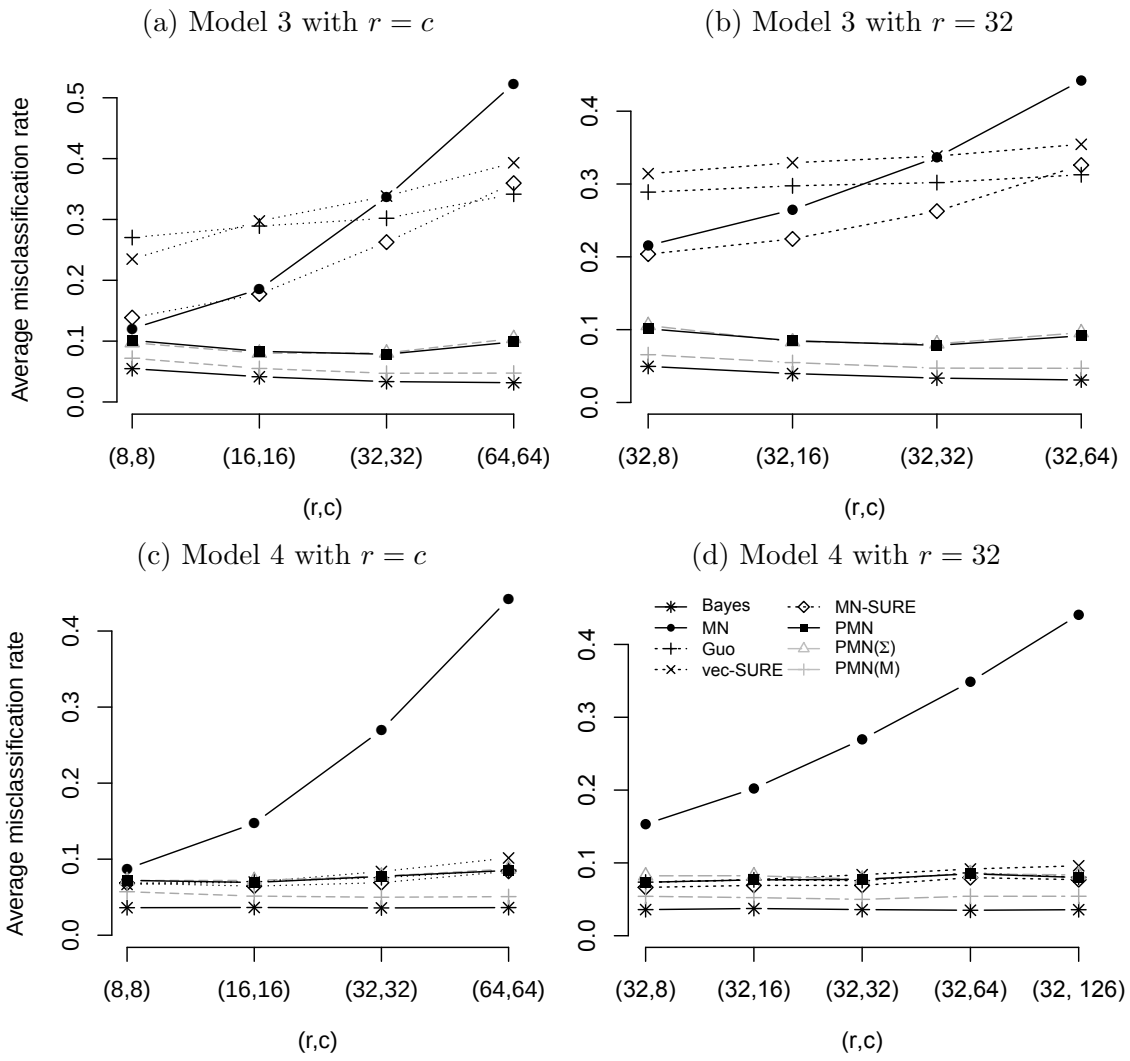



Figure 2.3: Misclassification rates averaged over 100 replications; (a) and (b) are for Model 3 and (c) and (d) for Model 4.

(a) Model 1 with $(r, c) = (32, 32)$ (b) Model 1 with $(r, c) = (64, 64)$

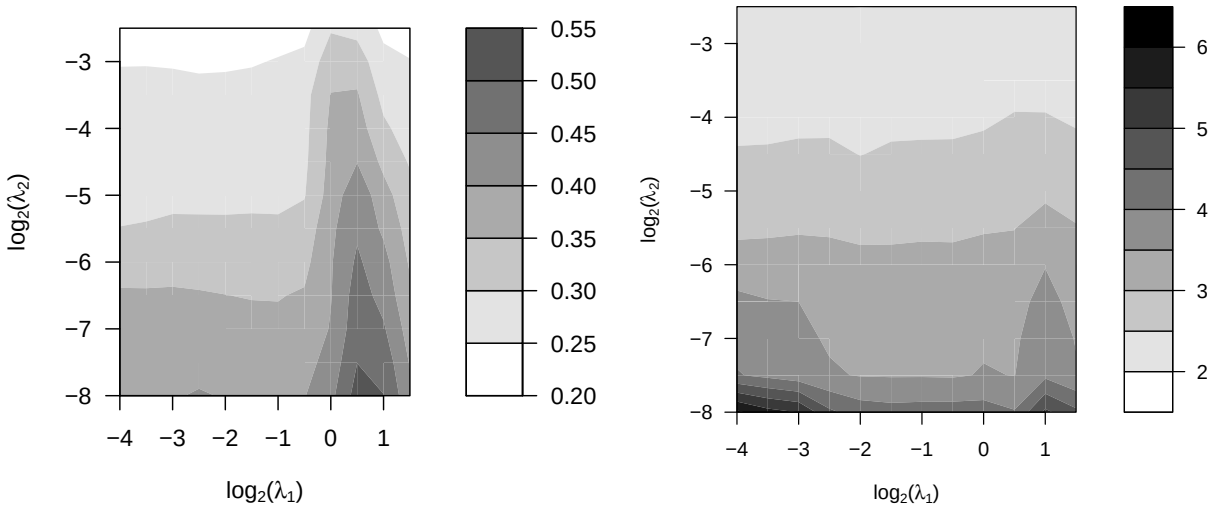



Figure 2.4: Smoothed contour plots of average computing times (in seconds) over 100 replications for each of $12 \times 12$ candidate grid points under Model 1 using warm-starts.

Table 2.2: Summary statistics for average computing time (in seconds) over 100 replications for PMN under Model 1. Candidate grid timings show the minimum, median, mean, and maximum average computing time over a $12 \times 12$ grid of candidate tuning parameters using warm-starts. The columns corresponding to $(\hat{\lambda}_1, \hat{\lambda}_2)$ give the average computing time (without warm-start initialization) for the tuning parameter pair chosen to minimize the misclassification rate on the validation set.

| | Model 1 $r = c$ | | | | | | Model 1 $r = 32$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Candidate Grid | | | | $(\hat{\lambda}_1, \hat{\lambda}_2)$ | | Candidate Grid | | | | $(\hat{\lambda}_1, \hat{\lambda}_2)$ | |
| | Min | Median | Mean | Max | Median | Mean | Min | Median | Mean | Max | Median | Mean |
| $c = 8$ | 0.019 | 0.027 | 0.030 | 0.056 | 0.076 | 0.070 | 0.044 | 0.073 | 0.080 | 0.159 | 0.189 | 0.177 |
| $c = 16$ | 0.045 | 0.068 | 0.073 | 0.122 | 0.178 | 0.162 | 0.077 | 0.126 | 0.131 | 0.244 | 0.295 | 0.295 |
| $c = 32$ | 0.212 | 0.332 | 0.332 | 0.521 | 0.608 | 0.822 | 0.212 | 0.332 | 0.332 | 0.521 | 0.608 | 0.822 |
| $c = 64$ | 1.991 | 2.846 | 2.991 | 6.092 | 4.336 | 4.917 | 0.752 | 1.130 | 1.605 | 6.355 | 1.849 | 1.815 |
| $c = 126$ | – | – | – | – | – | – | 2.978 | 3.340 | 40.985 | 245.173 | 2.061 | 12.367 |

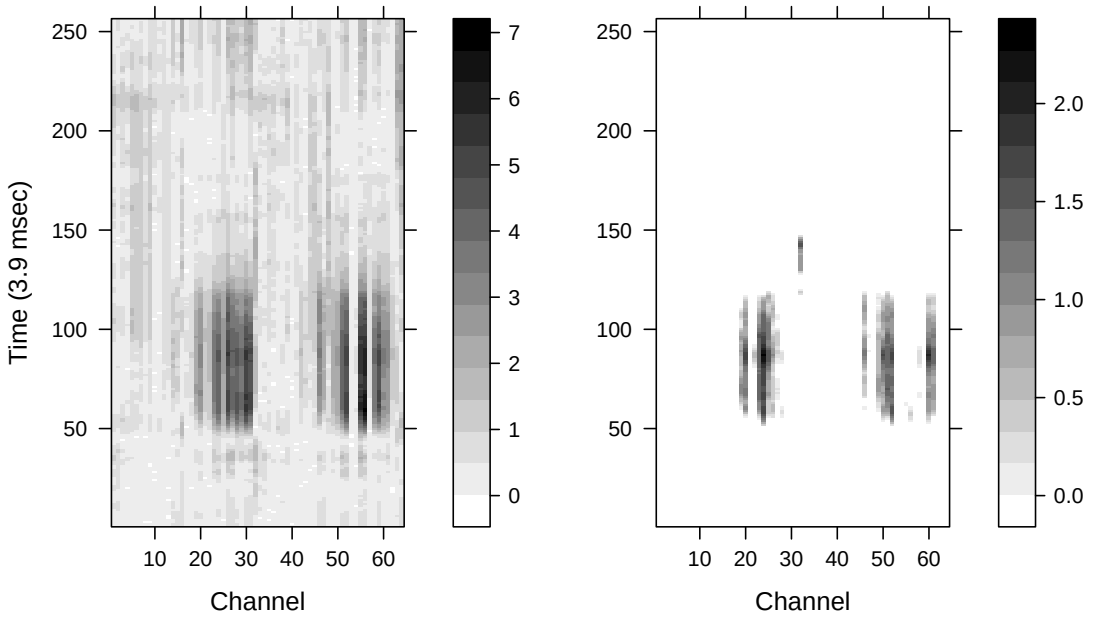
## 2.5 EEG data example

We analyzed the EEG data (https://kdd.ics.uci.edu/databases/eeg/eeg.html) also studied by Li et al. (2010) and Zhou and Li (2014). In the original study, 122 subjects, 77 of whom were alcoholics and 45 of whom were control, were exposed to stimuli while voltage was measured from $c = 64$ channels on a subject's scalp at $r = 256$ time points. Each subject underwent 120 trials. Each trial had one of three possible stimuli: single stimulus, two matched stimuli, or two unmatched stimuli. As in Li et al. (2010) and Zhou and Li (2014), we only analyze the single stimulus condition. Because each subject underwent multiple trials under the single stimulus condition, we use the within subject average over all single stimulus trials as the predictor and we use whether they were alcoholic or control as the response.

It is common to assume that (2.2) holds in the analysis of EEG data. For example, Zhou (2014) assumed that (2.2) holds when analyzing a single subject from this same dataset. It may also be reasonable to assume that only a subset of channels and time point combinations are important for discriminating between alcoholic and control response categories. Thus, the primary goal of our analysis is to identify a subset of channels and time point combinations that help explain how the alcoholics and controls react to the stimulus differently.

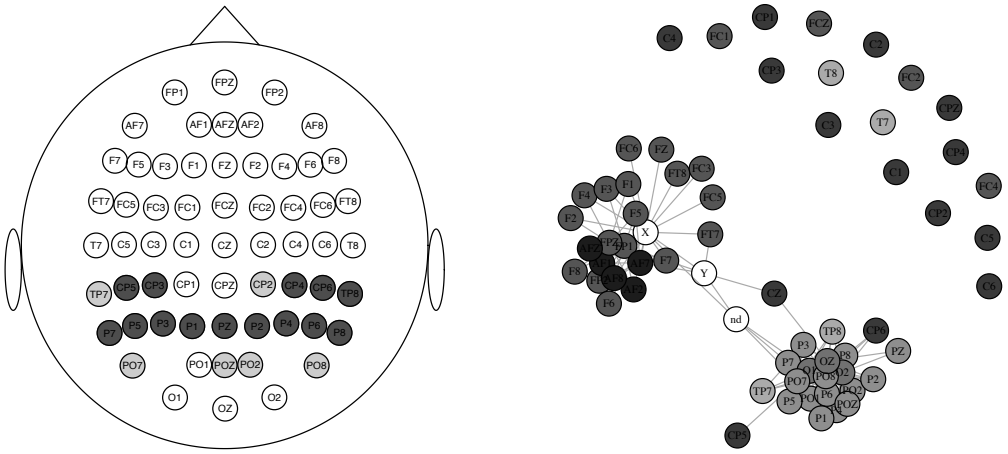
To demonstrate our method's classification accuracy, we used the leave-one-out cross validation approach from Li et al. (2010) and Zhou and Li (2014). For $k = 1, \ldots, 122$, we left out the $k$th observation and used the remaining 121 observations as training data. For each $k$, we selected tuning parameters for use in (2.3) by minimizing 5-fold cross validation misclassification error on the training dataset. Our method correctly classified 97 of 122 observations. Li et al. (2010) and Zhou and Li (2014) reported correctly classifying 97 and 94 of 122, respectively. Li et al. (2010) used quadratic discriminant analysis after dimension-folding of the predictors, and Zhou and Li (2014) used logistic regression with spectral regularization of the coefficient matrix.

To demonstrate the interpretability our fitted model, we separately fit (2.3) using the




(a) (b)

Figure 2.5: (a) The absolute value of the sample mean differences between the alcoholic and control response categories. (b) The absolute value of the estimated mean differences from (2.3) based on the tuning parameter pair $(\lambda_1, \lambda_2) = (0.15, 5.66)$, which had leave-one-out cross-validation classification accuracy of 98 out of 122.




(a) (b)

Figure 2.6: (a) An EEG cap based on the fitted model using $(\lambda_1, \lambda_2) = (0.15, 5.66)$. Dark grey channels had at least twenty time points estimated to have nonzero mean differences; light grey channels had less than twenty but greater than zero, whereas white channels had no nonzero mean differences. (b) The Gaussian precision graphical model corresponding to $\hat{\Delta}$. Different shades of grey correspond to different regions of the EEG channels; white channels are those that do not appear on the EEG cap image.

complete dataset. We used the tuning parameter pair $(\lambda_1, \lambda_2) = (0.15, 5.66)$, which had leave-one-out classification accuracy of 98 out of 122. The estimated mean difference, displayed as a heatmap in Figure 2.5(b), had 15466 of 16384 entries equal to zero.

Our fitted model can be used to easily identify which channels and time points have nonzero mean differences. We estimated only 22 of the 64 channels to have at least one time point where the mean differences were nonzero, only 16 of which had at least 20 nonzero time points. Inspecting the estimated mean differences displayed in Figure 2.5, it seems that the majority of activity that distinguishes between the alcoholic and control subjects takes place between the 52nd and 115th time points. We used the `R` package `eegkit` (Helwig, 2015) to display which channels had nonzero mean differences in Figure 2.5a. Our method does not explicitly use the spatial structure of channels in estimation, yet it recovered a set of important channels which have a natural arrangement in space.

Both $\Phi_*$ and $\Delta_*$ were estimated to be relatively sparse: $\hat{\Phi}$ was a diagonal matrix, while

$\hat{\Delta}$ had 3676 of 4032 off-diagonals equal to zero. Our estimate $\hat{\Delta}$ can be interpreted in terms of a Gaussian precision graphical model corresponding to the conditional dependence structure of the channels. We display the graphical model corresponding to $\hat{\Delta}$ in Figure 2.6(b). The graph structure corresponds to the spatial arrangement of channels displayed in Figure 2.6(a) – a result also observed by Zhou (2014).

## 2.6 Extensions

Our method naturally extends to the quadratic discriminant analysis model, where one assumes

$$\text{vec}(X \mid Y = j) \sim \text{N}_{rc}\left\{\text{vec}\left(\mu_{*j}\right), \Sigma_{*j}\right\},$$

where $\Sigma_{*j} \in \mathbb{S}^+_{rc}$ is the covariance matrix for the $j$th class for all $j \in \mathcal{J}$. To generalize (2.2), one can assume either (i) $\Sigma_{*j} = \Delta_{*j} \otimes \Phi_{*j}$, (ii) $\Sigma_{*j} = \Delta_{*j} \otimes \Phi_{*}$, and (iii) $\Sigma_{*j} = \Delta_{*} \otimes \Phi_{*j}$, where under (ii) and (iii), only one of the two component precision matrices are unique to each response category. Our algorithms can be easily modified to accommodate (i),(ii), or (iii).

The assumption (2.2) and estimator (2.3) can be generalized to cases where the predictor is a multidimensional array of order three or more, such as in fMRI or video data. In this case, where $x_i \in \mathbb{R}^{d_1 \times \cdots \times d_q}$, we can generalize the assumption (2.2) to the matrix $\Sigma_* \in \mathbb{S}^K_+$ where $K = \prod_{i=1}^q d_i$ so that

$$\Sigma_*^{-1} = \Xi_1 \otimes \cdots \otimes \Xi_q, \tag{2.14}$$

where $\Xi_l \in \mathbb{S}^+_{d_l}$ for $l = 1, \ldots, q$. Under (2.14), (2.1) becomes the array-normal distribution (Hoff, 2011; Manceur and Dutilleul, 2013). Algorithm 1 can be generalized to this setting using the same arguments as in Section 2.3. In particular, our alternating minimization algorithm could be applied by replacing the matrix product in the right hand side of (2.11) with a tensor product. Special computational considerations may be necessary when $\max\{d_1, \ldots, d_q\}$ is large. For instance, it may be better to approximate (2.3) in two steps;

the first step would estimate the covariance parameters with $\mu_j$ fixed at $\bar{x}_j$ for $j = 1, \ldots, J$, and the second step would estimate $\mu_*$ using (2.7) with the precision matrix components fixed at their estimates.

# Chapter 3

# Indirect multivariate response linear regression

## 3.1 Introduction

Some statistical applications require the modeling of a multivariate response. Let $y_i \in \mathbb{R}^q$ be the measurement of the $q$-variate response for the $i$th subject and let $x_i \in \mathbb{R}^p$ be the nonrandom values of the $p$ predictors for the $i$th subject ($i = 1, \ldots, n$). The multivariate response linear regression model assumes that $y_i$ is a realization of the random vector

$$Y_i = \mu_* + \beta_*^{\mathrm{T}} x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where $\mu_* \in \mathbb{R}^q$ is the unknown intercept, $\beta_*$ is the unknown $p$ by $q$ regression coefficient matrix, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent copies of a mean zero random vector with covariance matrix $\Sigma_{*E}$.

The ordinary least squares estimator of $\beta_*$ is

$$\hat{\beta}^{\mathrm{OLS}} = \underset{\beta \in \mathbb{R}^{p \times q}}{\arg\min} \|\mathbb{Y} - \mathbb{X}\beta\|_F^2, \tag{3.2}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\mathbb{R}^{p \times q}$ is the set of real valued $p$ by $q$ matrices, $\mathbb{Y}$ is the $n$ by $q$ matrix with $i$th row $(Y_i - n^{-1}\sum_{i=1}^n Y_i)^{\mathrm{T}}$, and $\mathbb{X}$ is the $n$ by $p$ matrix with $i$th row $(x_i - n^{-1}\sum_{i=1}^n x_i)^{\mathrm{T}}$ $(i = 1, ..., n)$. It is well known that $\hat{\beta}^{\mathrm{OLS}}$ is the maximum likelihood

estimator of $\beta_*$ when $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed $\mathrm{N}_q(0, \Sigma_{*E})$ and the corresponding maximum likelihood estimator of $\Sigma_{*E}^{-1}$ exists.

Many shrinkage estimators of $\beta_*$ have been proposed by penalizing the optimization in (3.2). Some simultaneously estimate $\beta_*$ and remove irrelevant predictors (Turlach et al., 2005; Obozinski et al., 2010; Peng et al., 2010). Others encourage an estimator of reduced rank (Yuan et al., 2007; Chen and Huang, 2012).

Under the restriction that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed $\mathrm{N}_q(0, \Sigma_{*E})$, shrinkage estimators of $\beta_*$ that penalize or constrain the minimization of the negative log-likelihood have been proposed. These methods simultaneously estimate $\beta_*$ and $\Sigma_{*E}^{-1}$. Examples include maximum likelihood reduced-rank regression (Izenman, 1975; Reinsel and Velu, 1998), envelope models (Cook et al., 2010; Su and Cook, 2011, 2012, 2013), and multivariate regression with covariance estimation (Rothman et al., 2010; Lee and Liu, 2012; Bhadra and Mallick, 2013).

To fit (3.1) with these shrinkage estimators, one exploits explicit assumptions about $\beta_*$ that may be unreasonable in some applications. As an alternative, we propose an indirect method to fit (3.1) without such assumptions. We instead assume that response and predictors have a joint multivariate normal distribution and we employ shrinkage estimators of the parameters of the conditional distribution of the predictors given the response. Our method provides alternative indirect estimators of $\beta_*$, which may be suitable when existing shrinkage estimators are inadequate. In the very challenging case when $p$ is large and $\beta_*$ is not sparse, one of our proposed indirect estimators can predict well and be interpreted easily.

## 3.2 A new class of indirect estimators of $\beta_*$

### 3.2.1 Class definition

We assume that the measured predictor and response pairs $(x_1, y_1), \ldots, (x_n, y_n)$ are a realization of $n$ independent copies of $(X, Y)$, where $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}} \sim \mathrm{N}_{p+q}(\mu_*, \Sigma_*)$. We also assume that $\Sigma_*$ is positive definite. Define the marginal parameters through the following

partitions:

$$\mu_* = \begin{pmatrix} \mu_{*X} \\ \mu_{*Y} \end{pmatrix}, \quad \Sigma_* = \begin{pmatrix} \Sigma_{*XX} & \Sigma_{*XY} \\ \Sigma_{*XY}^{\mathrm{T}} & \Sigma_{*YY} \end{pmatrix}.$$

Our goal is to estimate the multivariate regression coefficient matrix $\beta_* = \Sigma_{*XX}^{-1}\Sigma_{*XY}$ in the forward regression model

$$Y \mid X = x \sim \mathrm{N}_q\left\{\mu_{*Y} + \beta_*^{\mathrm{T}}\left(x - \mu_{*X}\right), \Sigma_{*E}\right\},$$

without assuming that $\beta_*$ is sparse or that $\|\beta_*\|_F^2$ is small. To do this we will estimate the inverse regression's coefficient matrix $\eta_* = \Sigma_{*YY}^{-1}\Sigma_{*XY}^{\mathrm{T}}$ and the inverse regression's error precision matrix $\Delta_*^{-1}$ in the inverse regression model

$$X \mid Y = y \sim \mathrm{N}_p\left\{\mu_{*X} + \eta_*^{\mathrm{T}}\left(y - \mu_{*Y}\right), \Delta_*\right\}.$$

We connect the parameters of the inverse regression model to $\beta_*$ with the following proposition, which we prove in Appendix A.1.

**Proposition 1**

*If $\Sigma_*$ is positive definite, then*

$$\beta_* = \Delta_*^{-1}\eta_*^{\mathrm{T}}\left(\Sigma_{*YY}^{-1} + \eta_*\Delta_*^{-1}\eta_*^{\mathrm{T}}\right)^{-1}. \tag{3.3}$$

This result leads us to propose a class of estimators of $\beta_*$,

$$\hat{\beta} = \hat{\Delta}^{-1}\hat{\eta}^{\mathrm{T}}(\hat{\Sigma}_{YY}^{-1} + \hat{\eta}\hat{\Delta}^{-1}\hat{\eta}^{\mathrm{T}})^{-1}, \tag{3.4}$$

where $\hat{\eta}$, $\hat{\Delta}$, and $\hat{\Sigma}_{YY}$ are user-selected estimators of $\eta_*$, $\Delta_*$, and $\Sigma_{*YY}$. If $n > p+q+1$ and the ordinary sample estimators are used for $\hat{\eta}$, $\hat{\Delta}$ and $\hat{\Sigma}_{YY}$, then $\hat{\beta}$ is equivalent to $\hat{\beta}^{\mathrm{OLS}}$.

We propose to use shrinkage estimators of $\eta_*$, $\Delta_*^{-1}$, and $\Sigma_{*YY}^{-1}$ in (3.4). This gives us the potential to indirectly fit an unparsimonious forward regression model by fitting a parsimonious inverse regression model. For example, $\eta_*$ could have only one nonzero entry

and all entries in $\beta_*$ could be nonzero. Conveniently, entries in $\eta_*$ can be interpreted like entries in $\beta_*$ are by reversing the roles of the predictors and responses. To fit the inverse regression model, we could use any of the forward regression shrinkage estimators discussed in Section 3.1.

### 3.2.2 Related work

Lee and Liu (2012) proposed an estimator of $\beta_*$ that also exploits the assumption that $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$ is multivariate normal; however, unlike our approach which makes no explicit assumptions about $\beta_*$, they assume that both $\Sigma_*^{-1}$ and $\beta_*$ are sparse.

Modeling the inverse regression is a well-known idea in multivariate analysis. For example, when $Y$ is categorical, quadratic discriminant analysis treats $X \mid Y = y$ as $p$-variate normal. There are also many examples of modeling the inverse regression in the sufficient dimension reduction literature (Adragni and Cook, 2009).

The work most closely related to ours is Cook et al. (2013). They proposed indirect estimators of $\beta_*$ based on modeling the inverse regression in the special case when the response is univariate, i.e., $q = 1$. Under our multivariate normal assumption on $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$, Cook et al. (2013) showed that

$$\beta_* = \frac{1}{1 + \Sigma_{*XY}^{\mathrm{T}} \Delta_*^{-1} \Sigma_{*XY} / \Sigma_{*YY}} \Delta_*^{-1} \Sigma_{*XY}, \tag{3.5}$$

and proposed estimators of $\beta_*$ by replacing $\Sigma_{*XY}$ and $\Sigma_{*YY}$ in (3.5) with their usual sample estimators, and by replacing $\Delta_*^{-1}$ with a shrinkage estimator. This class of estimators was designed to exploit an abundant signal rate in the forward univariate response regression when $p > n$.

## 3.3 Asymptotic analysis

We present a convergence rate bound for the indirect estimator of $\beta_*$ defined by (3.4). Our bound allows $p$ and $q$ to grow with the sample size $n$. In the following proposition, $\|\cdot\|$ is

the spectral norm and $\varphi_{\min}(\cdot)$ is the minimum eigenvalue.

**Proposition 2**

*Suppose that the following conditions are true: (i) $\Sigma_*$ is positive definite for all $p+q$; (ii) the estimator $\hat{\Sigma}_{YY}^{-1}$ is positive definite for all $q$; (iii) the estimator $\hat{\Delta}^{-1}$ is positive definite for all $p$; (iv) there exists a positive constant $K$ such that $\varphi_{\min}(\Sigma_{*YY}^{-1}) \geq K$ for all $q$; and (v) there exist sequences $\{a_n\}, \{b_n\}$ and $\{c_n\}$ such that $\|\hat{\eta} - \eta_*\| = O_P(a_n)$, $\|\hat{\Delta}^{-1} - \Delta_*^{-1}\| = O_P(b_n)$, $\|\hat{\Sigma}_{YY}^{-1} - \Sigma_{*YY}^{-1}\| = O_P(c_n)$, and $a_n\|\eta_*\|\|\Delta_*^{-1}\| + b_n\|\eta_*\|^2 + c_n \to 0$ as $n \to \infty$. Then*

$$\|\hat{\beta} - \beta_*\| = O_P\left(a_n\|\eta_*\|^2\|\Delta_*^{-1}\|^2 + b_n\|\eta_*\|^3\|\Delta_*^{-1}\| + c_n\|\eta_*\|\|\Delta_*^{-1}\|\right).$$

We prove Proposition 2 in Appendix A.1. We used the spectral norm because it is compatible with the convergence rate bounds established for sparse inverse covariance estimators (Rothman et al., 2008; Lam and Fan, 2009; Ravikumar et al., 2011).

If the inverse regression is parsimonious in the sense that $\|\eta_*\|$ and $\|\Delta_*^{-1}\|$ are bounded, then the bound in Proposition 2 simplifies to $\|\hat{\beta} - \beta_*\| = O_P(a_n + b_n + c_n)$. We explore finite-sample performance in Section 3.5.

## 3.4 Estimators in our class

### 3.4.1 Sparse inverse regression

We now describe an estimator of the forward regression coefficient matrix $\beta_*$ defined by (3.4) that exploits zeros in the inverse regression's coefficient matrix $\eta_*$, zeros in the inverse regression's error precision matrix $\Delta_*^{-1}$, and zeros in the precision matrix of the responses $\Sigma_{*YY}^{-1}$. We estimate $\eta_*$ with

$$\hat{\eta}^{\mathrm{L1}} = \underset{\eta \in \mathbb{R}^{q \times p}}{\arg\min} \left\{ \|\mathbb{X} - \mathbb{Y}\eta\|_F^2 + \sum_{j=1}^{p} \lambda_j \sum_{m=1}^{q} |\eta_{mj}| \right\}, \tag{3.6}$$

which separates into $p$ $L_1$-penalized least-squares regressions (Tibshirani, 1996): the first predictor regressed on the response through the $p$th predictor regressed on the response. We select $\lambda_j$ with 5-fold cross-validation, minimizing squared prediction error totaled over the folds, in the regression of the $j$th predictor on the response ($j = 1, \ldots, p$). This allows us to estimate the columns of $\eta_*$ in parallel.

We estimate $\Delta_*^{-1}$ and $\Sigma_{*YY}^{-1}$ with $L_1$-penalized Gaussian likelihood precision matrix estimation (Yuan and Lin, 2007; Banerjee et al., 2008). Let $\hat{\Sigma}_{\gamma,S}^{-1}$ be a generic version of this estimator with tuning parameter $\gamma$ and input $p$ by $p$ sample covariance matrix $S$:

$$\hat{\Sigma}_{\gamma,S}^{-1} = \underset{\Omega \in \mathbb{S}_+^p}{\arg\min} \left\{ \operatorname{tr}(\Omega S) - \log|\Omega| + \gamma \sum_{j \neq k} |\omega_{jk}| \right\}, \tag{3.7}$$

where $\mathbb{S}_+^p$ is the set of symmetric and positive definite $p$ by $p$ matrices. The optimization in (3.7) was used to estimate the inverse regression's error precision matrix in the univariate response regression methods proposed by Cook et al. (2012) and Cook et al. (2013). There are many algorithms that solve (3.7). Two good choices are the graphical lasso (Yuan, 2008; Friedman et al., 2008) and the algorithm of Hsieh et al. (2011). We select $\gamma$ with 5-fold cross-validation maximizing a validation likelihood criterion (Huang et al., 2006):

$$\hat{\gamma} = \underset{\gamma \in \mathcal{G}}{\arg\min} \sum_{k=1}^{5} \left\{ \operatorname{tr}\left( \hat{\Sigma}_{\gamma,S_{(-k)}}^{-1} S_{(k)} \right) - \log\left| \hat{\Sigma}_{\gamma,S_{(-k)}}^{-1} \right| \right\}, \tag{3.8}$$

where $\mathcal{G}$ is a user-selected finite subset of the non-negative real line, $S_{(-k)}$ is the sample covariance matrix from the observations outside the $k$th fold, and $S_{(k)}$ is the sample covariance matrix from the observations in the $k$th fold centered by the sample mean of the observations outside the $k$th fold. We estimate $\Delta_*^{-1}$ using (3.7) with its tuning parameter selected by (3.8) and $S = (\mathbb{X} - \mathbb{Y}\hat{\eta}^{\mathrm{L1}})^{\mathrm{T}}(\mathbb{X} - \mathbb{Y}\hat{\eta}^{\mathrm{L1}})/n$. Similarly, we estimate $\Sigma_{*YY}^{-1}$ using (3.7) with its tuning parameter selected by (3.8) and $S = \mathbb{Y}^{\mathrm{T}}\mathbb{Y}/n$.

### 3.4.2 Reduced-rank inverse regression

We propose indirect estimators of $\beta_*$ that presupposes that the inverse regression's coefficient matrix $\eta_*$ is rank-deficient. The following proposition links rank deficiency in $\eta_*$ and its estimator to $\beta_*$ and its indirect estimator.

**Proposition 3**

*If* $\Sigma_*$ *is positive definite, then* $\operatorname{rank}(\beta_*) = \operatorname{rank}(\eta_*)$. *In addition, if* $\hat{\Sigma}_{YY}^{-1}$ *and* $\hat{\Delta}^{-1}$ *are positive definite in the indirect estimator* $\hat{\beta}$ *defined by* (3.4), *then* $\operatorname{rank}(\hat{\beta}) = \operatorname{rank}(\hat{\eta})$.

We propose two reduced-rank indirect estimators of $\beta_*$ by inserting estimators of $\eta_*, \Delta_*^{-1}$, and $\Sigma_{*YY}$ in (3.4). The first estimates $\Sigma_{*YY}$ with $\mathbb{Y}^{\mathrm{T}}\mathbb{Y}/n$ and estimates $(\eta_*, \Delta_*^{-1})$ with normal likelihood reduced-rank inverse regression:

$$(\hat{\eta}^{(r)}, \hat{\Delta}^{-1(r)}) = \underset{(\eta,\Omega)\in\mathbb{R}^{q\times p}\times\mathbb{S}_+^p}{\arg\min} \left[n^{-1}\operatorname{tr}\left\{(\mathbb{X}-\mathbb{Y}\eta)^{\mathrm{T}}(\mathbb{X}-\mathbb{Y}\eta)\Omega\right\} - \log\det(\Omega)\right] \tag{3.9}$$

$$\text{subject to } \operatorname{rank}(\eta) = r,$$

where $r$ is selected from $\{0, \ldots, \min(p, q)\}$. The solution to (3.9) is available in closed form (Reinsel and Velu, 1998).

The second reduced-rank indirect estimator of $\beta_*$ estimates $\eta_*$ with $\hat{\eta}^{(r)}$ defined in (3.9), estimates $\Sigma_{*YY}^{-1}$ with (3.7) using $S = \mathbb{Y}^{\mathrm{T}}\mathbb{Y}/n$, and estimates $\Delta_*^{-1}$ with (3.7) using $S = (\mathbb{X} - \mathbb{Y}\hat{\eta}^{(r)})^{\mathrm{T}}(\mathbb{X} - \mathbb{Y}\hat{\eta}^{(r)})/n$.

The first indirect estimator is likelihood-based and the second indirect estimator exploits sparsity in $\Sigma_{*YY}^{-1}$ and $\Delta_*^{-1}$. Neither estimator is defined when $\min(p, q) > n$. In this case, which we do not address, a regularized reduced-rank estimator of $\eta_*$ could be used instead of the estimator defined in (3.9), e.g., the factor estimation and selection estimator (Yuan et al., 2007) or the reduced-rank ridge regression estimator (Mukherjee and Zhu, 2011).

## 3.5 Simulations

### 3.5.1 Sparse inverse regression simulation

For 200 independent replications, we generated a realization of $n$ independent copies of $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$, where $Y \sim \mathrm{N}_q(0, \Sigma_{*YY})$ and $X \mid Y = y \sim \mathrm{N}_p(\eta_*^{\mathrm{T}} y, \Delta_*)$. The $(i,j)$th entry of $\Sigma_{*YY}$ was set to $\rho_Y^{|i-j|}$ and the $(i,j)$th entry of $\Delta_*$ was set to $\rho_\Delta^{|i-j|}$. We set $\eta_* = Z \circ A$, where $Z$ had entries independently drawn from $\mathrm{N}(0,1)$, $A$ had entries independently drawn from the Bernoulli distribution with nonzero probability $s_*$, and $\circ$ is the element-wise product. This model is ideal for the example estimator from Section 3.4.1 because $\Delta_*^{-1}$ and $\Sigma_{*YY}^{-1}$ are both sparse. In the settings we considered, every entry in the corresponding randomly generated $\beta_*$ is nonzero with high probability, but the magnitudes of these entries are small. This motivated us to compare our indirect estimators of $\beta_*$ to direct estimators of $\beta_*$ that use penalized least squares.

To evaluate performance, we used model error (Breiman and Friedman, 1997; Yuan et al., 2007), defined as

$$\mathrm{ME}(\hat{\beta}, \beta_*) = \mathrm{tr}\left\{(\hat{\beta} - \beta_*)^{\mathrm{T}} \Sigma_{*XX} (\hat{\beta} - \beta_*)\right\}. \tag{3.10}$$

In each replication, we recorded the observed model error for $I_1$, the indirect estimator proposed in Section 3.4.1; $I_S$, the indirect estimator defined by (3.4) with $\hat{\eta}$ defined by (3.6), $\hat{\Sigma}_{YY} = \mathbb{Y}^{\mathrm{T}}\mathbb{Y}/n$, and $\hat{\Delta} = (\mathbb{X} - \mathbb{Y}\hat{\eta}^{L1})^{\mathrm{T}}(\mathbb{X} - \mathbb{Y}\hat{\eta}^{L1})/n$; $O_\Delta$, a part-oracle indirect estimator defined by (3.4) with $\hat{\eta}$ defined by (3.6), $\hat{\Sigma}_{YY}^{-1}$ defined by (3.7), and $\hat{\Delta}^{-1} = \Delta_*^{-1}$; $O$, a part-oracle indirect estimator defined by (3.4) with $\hat{\eta}$ defined by (3.6), $\hat{\Sigma}_{YY}^{-1} = \Sigma_{*YY}^{-1}$, and $\hat{\Delta}^{-1} = \Delta_*^{-1}$; and $O_Y$, a part-oracle indirect estimator defined by (3.4) with $\hat{\eta}$ defined by (3.6), $\hat{\Sigma}_{YY}^{-1} = \Sigma_{*YY}^{-1}$, and $\hat{\Delta}^{-1}$ defined by (3.7). We also recorded the observed model error for the ordinary least squares estimator $(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{Y}$ when $n > p$; and the Moore–Penrose least squares estimator $\mathbb{X}^{-}\mathbb{Y}$, where $\mathbb{X}^{-}$ is the Moore–Penrose generalized inverse of $\mathbb{X}$ when $n \leq p$. In addition, we recorded the observed model error for the estimator formed by $q$ separate univariate ridge regressions, where tuning parameters were chosen separately; and

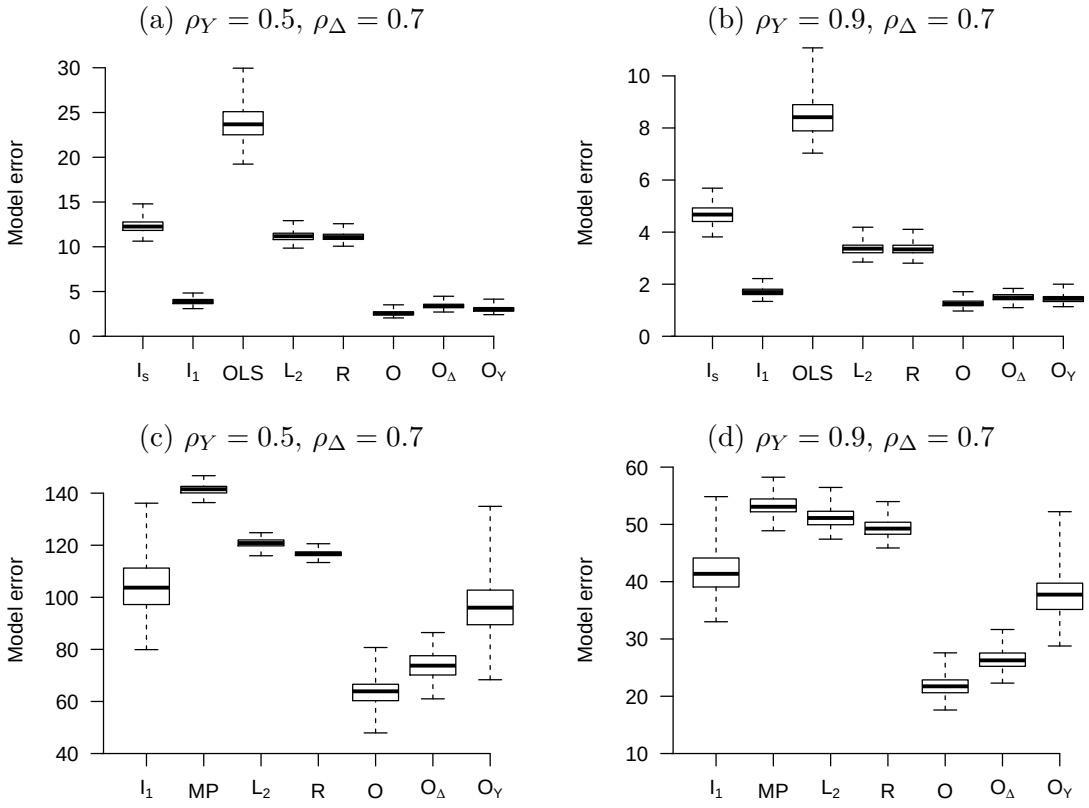



Figure 3.1: Boxplots of the observed model errors from 200 independent replications when the data generating model from Section 3.5.1 was used. In (a) and (b), $n = 100$, $p = 60$, $q = 60$, and $s_* = 0.1$. In (c) and (d), $n = 50$, $p = 200$, $q = 200$, and $s_* = 0.03$. The estimator OLS is ordinary least squares, MP is Moore–Penrose least squares, $L_2$ is $q$ univariate response ridge regressions with tuning parameters chosen separately, and R is multivariate ridge regression with one tuning parameter.

the multivariate ridge regression estimator, where a single tuning parameter was chosen.

We selected the tuning parameters for uses of (3.6) with 5-fold cross-validation, minimizing validation prediction error on the inverse regression. Tuning parameters for the ridge regression estimators were selected with 5-fold cross-validation, minimizing validation prediction error on the forward regression. We selected tuning parameters for uses of (3.7) with (3.8). The candidate set of tuning parameters was $\{10^{-8}, 10^{-7.5}, \ldots, 10^{7.5}, 10^{8}\}$.

We display side-by-side boxplots of the model errors from the 200 replications in Figure 3.1. When $n = 100$, $p = 60$, $q = 60$, and $s_* = 0.1$, the estimators based on (3.4)

performed well for both values of $\rho_Y$ that we considered. Our proposed estimator $I_1$ was even competitive with indirect estimators that used some oracle information. The version of our proposed estimator $I_s$ that used sample covariance matrices was outperformed by the forward regression estimators. This suggests that shrinkage estimation of $\Delta_*^{-1}$ and $\Sigma_{*YY}^{-1}$ was helpful.

When $n = 50$, $p = 200$, $q = 200$, and $s_* = 0.03$, our proposed indirect estimator $I_1$ outperformed all three forward regression estimators. The part-oracle method $O_\Delta$ that used the knowledge of $\Delta_*^{-1}$ outperformed the other part-oracle indirect estimator $O_Y$, which was slightly better than $I_1$. Additional results for this model are displayed in Appendix B. In those results, the performance of $I_1$ relative to the forward regression estimators was similar.

### 3.5.2 Non-normal forward regression simulation

For 200 independent replications, we generated $n$ independent copies of $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$ where $X \sim \mathrm{N}_p(0, \Sigma_{*XX})$ and $Y = \beta_*^{\mathrm{T}} X + \epsilon$. We set $\epsilon = \Sigma_{*E}^{1/2}(Z_1 - 1, \ldots, Z_q - 1)^{\mathrm{T}}$, where $Z_1, \ldots, Z_q$ are independent copies of an exponential random variable with mean 1. This ensures that $E(\epsilon) = 0$ and $\mathrm{Cov}(\epsilon) = \Sigma_{*E}$. We indirectly determined the entries of $\beta_*$, $\Sigma_{*E}$, and $\Sigma_{*XX}$ by specifying the entries in $\eta_*$, $\Delta_*^{-1}$, and $\Sigma_{*YY}$. This required us to use the multivariate normal model in Section 3.2.1 even though $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$ is not multivariate normal in this simulation. We set the $(i,j)$th entry in $\Sigma_{*YY}$ to $\rho_Y^{|i-j|}$ and the $(i,j)$th entry in $\Delta_*$ to $\rho_\Delta^{|i-j|}$. We also set $\eta_* = Z \circ A$, where $Z$ had entries independently drawn from $\mathrm{N}(0,1)$ and $A$ had entries independently drawn from the Bernoulli distribution with nonzero probability $s_*$. We compared the performance of the estimators described in Section 3.5.1 using model error. We selected tuning parameters in the same way that we did in the simulation described in Section 3.5.1.

We display side-by-side boxplots of the model errors from the 200 replications in Figure 3.2. The performance of $I_1$ relative to the competitors is similar to how it was in Section 3.5.1, where $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$ was multivariate normal.

We also performed simulations when $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$ had a multivariate elliptical $t$-distribution.

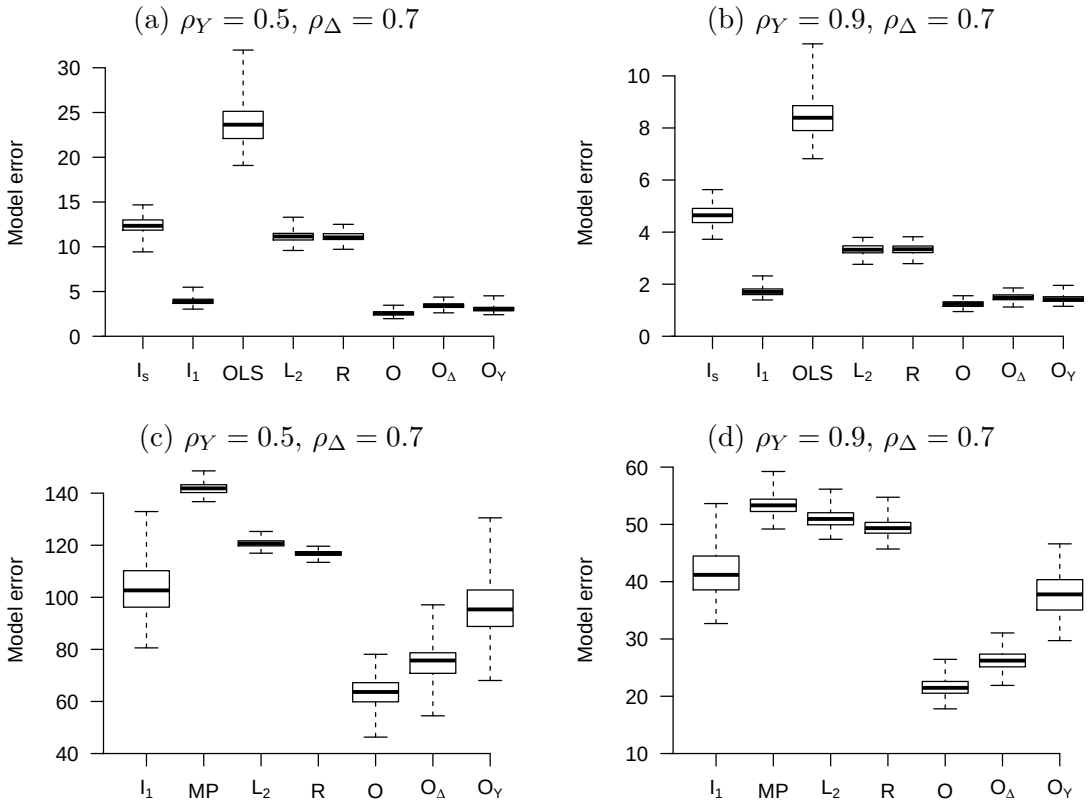



Figure 3.2: Boxplots of the observed model errors from 200 independent replications when the data generating model from Section 3.5.2 was used. In (a) and (b), $n = 100$, $p = 60$, $q = 60$, and $s_* = 0.1$. In (c) and (d), $n = 50$, $p = 200$, $q = 200$, and $s_* = 0.03$. The estimators are defined in Section 3.5.1 and the caption of Figure 3.1.

The results from this simulation are reported in Appendix B. When $n = 100$, $p = 60$, and $q = 60$, the results from the elliptical $t$-distribution simulation were similar to the results here. When $n = 50$, $p = 200$, $q = 200$ and the degrees of freedom of the elliptical $t$-distribution was small or the responses had weak marginal correlations, the proposed estimator $I_1$ was sometimes outperformed by competitors. These results suggest that our example estimator may work well for some non-normal data generating models.

### 3.5.3 Reduced-rank inverse regression simulation

For 200 independent replications, we generated a realization of $n$ independent copies of $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$ where $Y \sim \mathrm{N}_q(0, \Sigma_{*YY})$ and $X \mid Y = y \sim \mathrm{N}_p(\eta_*^{\mathrm{T}} y, \Delta_*)$. The $(i,j)$th entry of $\Sigma_{*YY}$ was set to $\rho_Y^{|i-j|}$ and the $(i,j)$th entry of $\Delta_*$ was set to $\rho_\Delta^{|i-j|}$. After specifying $r_* \leq \min(p,q)$, we set $\eta_* = PQ$, where $P \in \mathbb{R}^{q \times r_*}$ had entries independently drawn from $\mathrm{N}(0,1)$ and $Q \in \mathbb{R}^{r_* \times p}$ had entries independently drawn from $\mathrm{Uniform}(-0.25, 0.25)$ so that $r_* = \mathrm{rank}(\eta_*) = \mathrm{rank}(\beta_*)$.

In each replication, we measured the observed model error for $I_{\mathrm{ML}}$, the likelihood-based indirect first example estimator proposed in Section 3.4.2; $I_{\mathrm{RR}}$, the second indirect example estimator proposed in Section 3.4.2, which uses sparse estimators of $\Sigma_{*YY}^{-1}$ and $\Delta_*^{-1}$ in (3.4); $O_{R\Delta}$, a part-oracle indirect estimator defined by (3.4) with $\hat{\eta}$ defined by (3.9), $\hat{\Delta}^{-1}$ defined by (3.7), and $\hat{\Sigma}_{YY}^{-1} = \Sigma_{*YY}^{-1}$; $O_R$, a part-oracle indirect estimator defined by (3.4) with $\hat{\eta}$ defined by (3.9), $\hat{\Delta}^{-1} = \Delta_*^{-1}$, and $\hat{\Sigma}_{YY}^{-1} = \Sigma_{*YY}^{-1}$; $O_{RY}$, a part-oracle indirect estimator defined by (3.4) with $\hat{\eta}$ defined by (3.9), $\hat{\Delta}^{-1} = \Delta_*^{-1}$, $\hat{\Delta}^{-1}$ defined by (3.7), and $\hat{\Sigma}_{YY}^{-1}$ defined by (3.7). We also measured the observed model error for the direct likelihood-based reduced-rank regression estimator (Izenman, 1975; Reinsel and Velu, 1998) and the ordinary least squares estimator.

We selected the rank parameter $r$ for uses of (3.9) with 5-fold cross-validation, minimizing validation prediction error on the inverse regression. The rank parameter for the direct likelihood-based reduced-rank regression estimator was selected with 5-fold cross-validation, minimizing validation prediction error on the forward regression. We selected tuning parameters for uses of (3.7) with (3.8). The candidate set of tuning parameters was $\{10^{-8}, 10^{-7.5}, \ldots, 10^{7.5}, 10^{8}\}$.

We display side-by-side boxplots of the model errors for this reduced-rank inverse regression simulation in Figure 3.3 (a) and (b), where we set $n = 100$, $p = 20$, $q = 20$, and $r_* = 4$. This choice of $(n,p,q)$ ensures that $I_{\mathrm{ML}}$ exists with probability one. When $\rho_Y = 0.5$, $I_{\mathrm{RR}}$ outperformed all non-oracle competitors. When $\rho_Y = 0.9$, $I_{\mathrm{RR}}$ tended to outperform all non-oracle competitors, but it performed worse in a small number of replica-

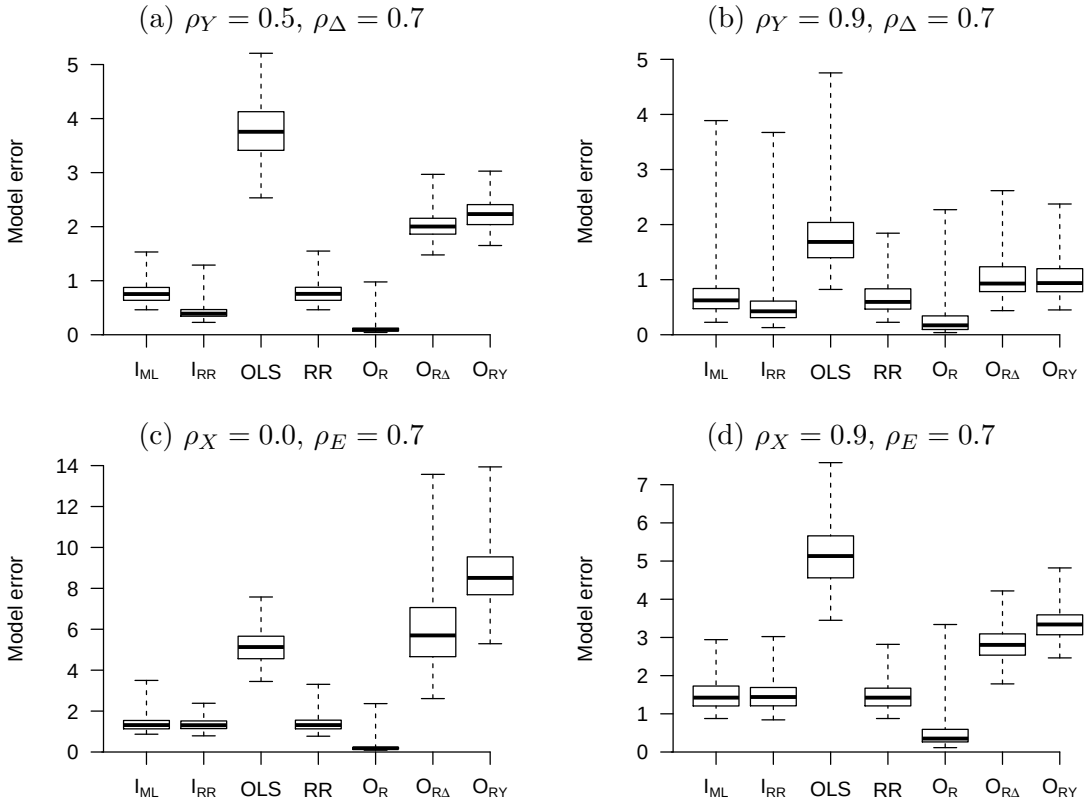



Figure 3.3: Boxplots of the observed model errors from 200 replications when $n = 100, p = 20$, $q = 20$, $r_* = 4$. In (a) and (b), the data generating model from Section 3.5.3 was used. In (c) and (d), the data generating model from Section 3.5.4 was used. The estimator RR is likelihood-based reduced-rank forward regression (Izenman, 1975; Reinsel and Velu, 1998) and OLS is ordinary least squares.

tions. Additionally, $I_{\mathrm{RR}}$ generally outperformed both $O_{R\Delta}$ and $O_{RY}$, which suggests that sparse estimation of $\Delta_*^{-1}$ and $\Sigma_{*YY}^{-1}$ was helpful. In each setting, $I_{\mathrm{ML}}$ performed similarly to the direct reduced-rank regression estimator even though they are estimating parameters of different conditional distributions. Simulation results from other data generating models are displayed in Appendix B.

### 3.5.4 Reduced-rank forward regression simulation

In this section, we compare the estimators from Section 3.5.3 using a forward regression data generating model.

For 200 independent replications, we generated a realization of $n$ independent copies of $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$ where $X \sim \mathrm{N}_p(0, \Sigma_{*XX})$ and $Y \mid X = x \sim \mathrm{N}_q(\beta_*^{\mathrm{T}} x, \Sigma_{*E})$. The $(i,j)$th entry of $\Sigma_{*XX}$ was set to $\rho_X^{|i-j|}$ and the $(i,j)$th entry of $\Sigma_{*E}$ was set to $\rho_E^{|i-j|}$. After specifying $r_* \leq \min(p,q)$, we set $\beta_* = ZQ$ where $Z \in \mathbb{R}^{p \times r_*}$ had entries independently drawn from $\mathrm{N}(0,1)$ and $Q \in \mathbb{R}^{r_* \times q}$ had entries independently drawn from $\mathrm{Uniform}(-0.25, 0.25)$. In this data generating model, neither $\Delta_*^{-1}$ nor $\Sigma_{*YY}^{-1}$ had entries equal to zero.

In each replication, we recorded the observed model error for the estimators described in Section 3.5.3. We present boxplots of these model errors from 200 replications with $n = 100$, $p = 20$, $q = 20$, and $r_* = 4$ in Figure 3.3 (c) and (d). Both $I_{\mathrm{RR}}$ and $I_{\mathrm{ML}}$ were competitive with the direct reduced-rank regression estimator. Although neither $\Delta_*^{-1}$ nor $\Sigma_{*YY}^{-1}$ were sparse, $I_{\mathrm{RR}}$ generally outperformed $O_{RY}$ and $O_{R\Delta}$, both of which use some oracle information. These results demonstrate that using sparse estimators of $\Delta_*^{-1}$ and $\Sigma_{*YY}^{-1}$ in (3.4) may be helpful when neither is truly sparse.

## 3.6 Genomic data example

We consider a comparative genomic hybridization dataset from Chin et al. (2006) analyzed by Witten et al. (2009) and Chen et al. (2013). The data are measured gene expression profiles and DNA copy-number variations for $n = 89$ subjects with breast cancer. We performed a separate multivariate response regression analysis for chromosomes 8, 17, and

22. In each analysis, the $q$-variate response was DNA copy-number variations and the $p$-variate predictor was the gene expression profile. The dimensions for the three analyses were $(p, q) = (673, 138)$, $(1161, 87)$, and $(618, 18)$.

In the analysis of Chen et al. (2013), estimators that used all $p$ genes significantly outperformed estimators that used a selected subset of genes. This may indicate that the forward regression coefficient matrix is not sparse. When analyzing similar data, Peng et al. (2010) and Yuan et al. (2012) focused on modeling the inverse regression, which they assumed to be sparse. This motivated us to apply our indirect estimator that also assumes that the inverse regression is sparse.

In each of 1000 replications, we randomly split the data into training and testing sets of sizes 60 and 29, respectively. Within each replication, we standardized the training dataset predictors and responses for model fitting and appropriately rescaled predictions. We fit the multivariate response linear regression model to the training dataset by estimating the regression coefficient matrix with non-oracle direct and indirect estimators described in Section 3.5.1. We modified our proposed estimator $I_1$ because computing the sparse estimates of $\Delta_*^{-1}$ and $\Sigma_{YY}^{-1}$ took too much time for small values of their tuning parameters. We instead used $I_2$, which is the same as $I_1$ except that the sparse estimators of $\Delta_*^{-1}$ and $\Sigma_{YY}^{-1}$ are replaced by the shrinkage estimator defined by

$$\underset{\Omega \in \mathbb{S}_+^p}{\operatorname{argmin}} \left\{ \operatorname{tr}(\Omega S) - \log \det(\Omega) + \gamma \sum_{j,k} |\omega_{jk}|^2 \right\}, \tag{3.11}$$

where $S = (\mathbb{Y} - \mathbb{X}\hat{\eta}^{L1})^{\mathrm{T}}(\mathbb{Y} - \mathbb{X}\hat{\eta}^{L1})/n$ when we estimate $\Delta_*^{-1}$, and $S = \mathbb{Y}^{\mathrm{T}}\mathbb{Y}/n$ when we estimate $\Sigma_{*YY}^{-1}$. Witten and Tibshirani (2009) derived a closed form solution for (3.11). This shrinkage estimator of the inverse regression's error precision matrix was also used in the data example of Cook et al. (2013). Tuning parameters were selected using the same procedures described in the simulation studies of Section 3.5, except the tuning parameter for $\hat{\Delta}^{-1}$ was chosen to minimize 5-fold cross-validation prediction error on the forward regression after having fixed $\hat{\eta}$ and $\hat{\Sigma}_{YY}^{-1}$. We also fit the model using the Moore–Penrose least squares estimator, $q$ separate lasso regressions, the multivariate group lasso estimator

of Obozinski et al. (2011), and both ridge regression estimators described in Section 3.5.

Tuning parameters for the direct estimators were chosen to minimize 5-fold cross-validation prediction error on the forward regression. In each replication, we measured the mean squared scaled prediction error which we define as

$$\frac{||(\mathbb{Y}_{\text{test}} - \mathbb{X}_{\text{test}}\hat{\beta})\Lambda^{-1}||_F^2}{29q},$$

where $\mathbb{Y}_{\text{test}} \in \mathbb{R}^{29\times q}$ is the test dataset response matrix column-centered by the training dataset response sample mean, $\mathbb{X}_{\text{test}} \in \mathbb{R}^{29\times p}$ is the test dataset predictor matrix column-centered by the training dataset predictor sample mean, and $\Lambda \in \mathbb{R}^{q\times q}$ is a diagonal matrix with the complete data response marginal standard deviations on its the diagonal. This measure puts predictions on the same scale for comparison across the $q$ responses.

The mean squared scaled prediction errors are summarized in Table 3.1. For all three chromosomes, the proposed estimator $I_2$ was better than the Moore–Penrose least square estimator, the null model, $q$ separate lasso regressions, and the group lasso estimator. Although the proposed estimator $I_2$ performed similarly to both ridge regression estimators, $I_2$ has the advantage of fitting an interpretable parsimonious inverse regression with an interesting biological interpretation. Figure 3.4 displays a heatmap representing how frequently each inverse regression coefficient was estimated to be nonzero with method $I_2$ in the 1000 replications for the analysis of Chromosome 17. The estimated inverse regression coefficient matrices were 3.18%, 4.05%, and 14.7% nonzero on average for the analyses of Chromosomes 8, 17, and 22 respectively.

Table 3.1: Mean squared scaled prediction error averaged over 1000 replications times 10 and corresponding standard errors times 10.

| Chromosome | $q$ | $p$ | $I_2$ | NM | MP | $L_1$ | $L_{1/2}$ | $L_2$ | R |
|---|---|---|---|---|---|---|---|---|---|
| 8 | 138 | 673 | 6.43 | 10.08 | 6.79 | 7.09 | 7.36 | 6.47 | 6.41 |
| | | | (0.029) | (0.052) | (0.029) | (0.033) | (0.035) | (0.030) | (0.030) |
| 17 | 87 | 1161 | 7.83 | 10.18 | 8.18 | 8.62 | 8.91 | 8.04 | 7.94 |
| | | | (0.046) | (0.064) | (0.046) | (0.049) | (0.050) | (0.050) | (0.049) |
| 22 | 18 | 618 | 6.05 | 10.37 | 6.67 | 6.86 | 6.62 | 6.15 | 6.13 |
| | | | (0.043) | (0.086) | (0.038) | (0.052) | (0.047) | (0.048) | (0.049) |

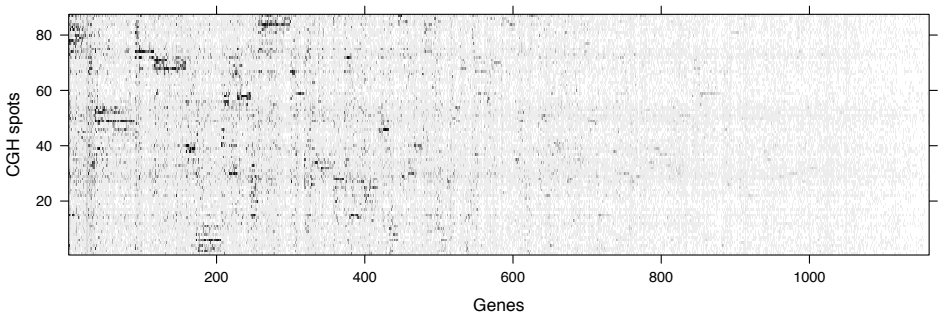



Figure 3.4: A heatmap displaying the number of replications out of 1000 for which entries in the inverse regression's coefficient matrix were estimated to be nonzero by $I_2$ for Chromosome 17. Black denotes 1000 and white denotes zero. The genes were sorted by hierarchical clustering.

## 3.7 Discussion

If one has access to the joint distribution of the predictors and responses, then one could use shrinkage estimators to fit both the forward and inverse regression models. One could then select the more parsimonious direction, which could be determined by the complexity of the models recommended by cross validation. If the inverse regression model is more parsimonious, then our method could be used to improve prediction in the forward direction. Prediction may be the only goal, in which case the forward and indirect predictions could be combined.

A referee for Molstad and Rothman (2016a) pointed out that it is expensive to compute an indirect estimator in our class when $q$ is very large because it requires the inversion of a $q$ by $q$ matrix in (3.4). This referee also mentioned that our class of indirect estimators is inapplicable when either the predictors or responses are categorical.

# Chapter 4

# Shrinking characteristics of precision matrix estimators

## 4.1 Introduction

Estimating precision matrices is required to fit many statistical models. Many papers written in the last decade have proposed shrinkage estimators of the precision matrix when $p$, the number of variables, is large. Pourahmadi (2013) and Fan et al. (2016) provide comprehensive reviews of large covariance and precision matrix estimation. The main strategy used in many of these papers is minimize the Gaussian negative log-likelihood plus a penalty on the off-diagonal entries of the optimization variable corresponding to the precision matrix. For example, Yuan and Lin (2007) proposed the $L_1$-penalized Gaussian likelihood precision matrix estimator defined by

$$\underset{\Omega \in \mathbb{S}_+^p}{\arg\min} \left\{ \operatorname{tr}(S\Omega) - \log\det(\Omega) + \lambda \sum_{i \neq j} |\Omega_{ij}| \right\}, \tag{4.1}$$

where $S$ is the sample covariance matrix, $\lambda > 0$ is a tuning parameter, $\mathbb{S}_+^p$ is the set of $p \times p$ symmetric and positive definite matrices, and tr and det are the trace and determinant, respectively. Other authors have replaced the $L_1$ penalty in (4.1) with the squared Frobenius norm (Witten and Tibshirani, 2009; Rothman and Forzani, 2014) or non-convex penalties that also encourage zeros in the estimator (Lam and Fan, 2009; Fan et al., 2009).

To fit many predictive models, only a characteristic of the population precision matrix needs to be estimated. For example, in binary linear discriminant analysis, the population precision matrix is needed for prediction only through the product of the precision matrix and the difference between the two conditional distribution mean vectors. Many authors have proposed methods that directly estimate this characteristic (Cai and Liu, 2011; Fan et al., 2012; Mai et al., 2012).

We propose to estimate the precision matrix by shrinking the characteristic of the estimator that is needed for prediction. The characteristic we consider is a linear or affine function evaluated at the precision matrix. The goal is to improve prediction performance. Unlike methods that estimate the characteristic directly, our approach provides the practitioner an estimate of the entire precision matrix, not just the characteristic. In our simulation studies and data example, we show that penalizing the characteristic needed for prediction can improve prediction performance over competing sparse precision estimators like (4.1), even when the true precision matrix is very sparse. In addition, estimators in our framework can be used in applications other than linear discriminant analysis.

## 4.2 Proposed method

### 4.2.1 Penalized likelihood estimator

We propose to estimate the population precision matrix $\Omega_*$ with

$$\hat{\Omega} = \underset{\Omega \in \mathbb{S}_+^p}{\arg\min} \left\{ \operatorname{tr}(S\Omega) - \log\det(\Omega) + \lambda |A\Omega B - C|_1 \right\}, \tag{4.2}$$

where $A \in \mathbb{R}^{a\times p}$, $B \in \mathbb{R}^{p\times b}$, and $C \in \mathbb{R}^{a\times b}$ are user-specified matrices; and $|M|_1 = \sum_{i,j} |M_{ij}|$. Our estimator exploits the assumption that $A\Omega_* B - C$ is sparse. In cases where $A$, $B$, and $C$ need to be estimated, we replace them with their estimators.

An estimator defined by (4.2) with $C = 0$ was mentioned in an unpublished manuscript by Dalal and Rajaratnam (2014) available on arXiv. These authors proposed an alternating minimization algorithm for solving (4.1) and described how to apply it to solve (4.2). Dalal

and Rajaratnam did not describe applications or theoretical properties of this estimator. Also, as written, their algorithm does not actually solve (4.2) when $A$ and $B$ are arbitrary. We propose an alternating direction method of multipliers algorithm to solve (4.2) and establish theoretical properties for this estimator.

### 4.2.2 Example applications

Fitting the discriminant analysis model requires the estimation of one or more precision matrices. In particular, the linear discriminant analysis model assumes that the data are independent copies of the random pair $(X, Y)$, where the support of $Y$ is $\{1, \ldots, J\}$ and

$$X \mid Y = j \sim \mathrm{N}_p\left(\mu_{*j}, \Omega_*^{-1}\right), \quad j = 1, \ldots, J, \tag{4.3}$$

where $\mu_{*j} \in \mathbb{R}^p$ and $\Omega_*^{-1} \in \mathbb{S}_+^p$ are unknown. To discriminate between response categories $l$ and $m$, only the characteristic $\Omega_*(\mu_{*l} - \mu_{*m})$ is needed. Methods that estimate this characteristic directly have been proposed (Cai and Liu, 2011; Mai et al., 2012; Fan et al., 2012; Mai et al., 2015). These methods are useful in high dimensions because they perform variable selection. For the $j$th variable to be non-informative for discriminating between response categories $l$ and $m$, it must be that the $j$th element of $\Omega_*(\mu_{*l} - \mu_{*m})$ is zero. While these methods can perform well in classification and variable selection, they do not actually fit the model in (4.3).

Methods for fitting (4.3) specifically for linear discriminant analysis either assume $\Omega_*$ is diagonal (Bickel and Levina, 2004) or that both $\mu_{*l} - \mu_{*m}$ and $\Omega_*$ are sparse (Guo, 2010; Xu et al., 2015). A method for fitting (4.3) and performing variable selection was proposed by Witten and Tibshirani (2009). They suggest a two-step procedure where one first estimates $\Omega_*$, and then with the estimate $\bar{\Omega}$ fixed, estimates each $\mu_{*j}$ by penalizing the characteristic $\bar{\Omega}\mu_j$, where $\mu_j$ is the optimization variable corresponding to $\mu_{*j}$.

To apply our method to the linear discriminant analysis problem, we use (4.2) with $A = I_p$, $C = 0$, and $B$ equal to the matrix whose columns are $\bar{x}_j - \bar{x}_k$ for all $1 \leq j < k \leq J$, where $\bar{x}_j$ is the unpenalized maximum likelihood estimator of $\mu_{*j}$. For large values of the

tuning parameter, this would lead an estimator of $\Omega_*$ such that $\hat{\Omega}(\bar{x}_j - \bar{x}_k)$ is sparse. Thus our approach simultaneously fits (4.3) and performs variable selection.

Precision and covariance matrix estimators are also needed for portfolio allocation. The optimal allocation based on the Markowitz (1952) minimum-variance portfolio is proportional to $\Omega_*\mu_*$, where $\mu_*$ is the vector of expected returns for $p$ assets and $\Omega_*$ is precision matrix for the returns. In practice, one would estimate $\Omega_*$ and $\mu_*$ with their usual sample estimators $\hat{\Omega}$ and $\hat{\mu}$. However, when $p$ is large, the usual sample estimator of $\Omega_*$ does not exist, so regularization is necessary. Moreover, Brodie et al. (2009) argue that sparse portfolios are often desirable when $p$ is large. While many have proposed using sparse or shrinkage estimators of $\Omega_*$ or $\Omega_*^{-1}$ plugged-in to the Markowitz criterion, e.g., Xue et al. (2012), this would not necessarily lead to sparse estimators of $\Omega_*\mu_*$. Chen et al. (2016) proposed a method for estimating the characteristic $\Omega_*\mu_*$ directly, but like the direct linear discriminant methods, this approach does not lead to an estimate of $\Omega_*$. For the sparse portfolio allocation problem, we propose to estimate $\Omega_*$ using (4.2) with $A = I_p$, $C = 0$, and $B = \hat{\mu}$.

Another application is in linear regression where the response and predictor have a joint multivariate normal distribution. In this case, the regression coefficient matrix is $\Omega_*\Sigma_{*XY}$, where $\Omega_*$ is the marginal precision matrix for the predictors and $\Sigma_{*XY}$ is the cross-covariance matrix between predictors and responses. We propose to estimate $\Omega_*$ using (4.2) with $A = I_p$, $C = 0$, and $B$ equal to the usual sample estimator of $\Sigma_{*XY}$. Similar to the procedure proposed by Witten and Tibshirani (2009), this approach provides an alternative method for estimating regression coefficients using shrinkage estimators of the marginal precision matrix for the predictors.

## 4.3 Computation

### 4.3.1 Alternating direction method of multipliers algorithm

To solve the optimization in (4.2), we propose an alternating direction method of multipliers algorithm with a modification based on the majorize-minimize principle (Lange, 2016).

Following the standard alternating direction method of multipliers approach (Boyd et al., 2011), we rewrite (4.2) as a constrained optimization problem:

$$\underset{(\Theta,\Omega)\in\mathbb{R}^{a\times b}\times\mathbb{S}_+^p}{\arg\min}\{\mathrm{tr}(S\Omega)-\log\det(\Omega)+\lambda|\Theta|_1\}\quad\text{subject to } A\Omega B-\Theta=C. \tag{4.4}$$

The augmented Lagrangian for (4.4) is defined by

$$\begin{aligned}\mathcal{F}_\rho(\Omega,\Theta,\Gamma)=\mathrm{tr}(S\Omega)-\log\det(\Omega)+\lambda|\Theta|_1\\ -\mathrm{tr}\left\{\Gamma^{\mathrm{T}}(A\Omega B-\Theta-C)\right\}+\frac{\rho}{2}\|A\Omega B-\Theta-C\|_F^2,\end{aligned}$$

where $\rho>0$ and $\Gamma\in\mathbb{R}^{a\times b}$ is the Lagrangian dual variable. Let the subscript $m$ denote the $m$th iterate. From Boyd et al. (2011), to solve (4.4), the alternating direction method of multipliers algorithm uses the following updating equations:

$$\Omega_{m+1}=\underset{\Omega\in\mathbb{S}_+^p}{\arg\min}\,\mathcal{F}_\rho(\Omega,\Theta_m,\Gamma_m), \tag{4.5}$$

$$\Theta_{m+1}=\underset{\Theta\in\mathbb{R}^{a\times b}}{\arg\min}\,\mathcal{F}_\rho(\Omega_{m+1},\Theta,\Gamma_m), \tag{4.6}$$

$$\Gamma_{m+1}=\Gamma_m-\rho\left(A\Omega_{m+1}B-\Theta_{m+1}-C\right). \tag{4.7}$$

Instead of solving (4.5) exactly, we approximate its objective function with a majorizing function. Specifically, we replace (4.5) with

$$\Omega_{m+1}=\underset{\Omega\in\mathbb{S}_p^+}{\arg\min}\left\{\mathcal{F}_\rho(\Omega,\Theta_m,\Gamma_m)+\frac{\rho}{2}\mathrm{vec}(\Omega-\Omega_m)^{\mathrm{T}}Q\mathrm{vec}(\Omega-\Omega_m)\right\}, \tag{4.8}$$

where $Q=\tau I-\left(A^{\mathrm{T}}A\otimes BB^{\mathrm{T}}\right)$, $\tau$ is selected so that $Q\in\mathbb{S}_+^p$, $\otimes$ is the Kronecker product, and vec forms a vector by stacking the columns of its matrix argument. Since

$$\mathrm{vec}(\Omega-\Omega_m)^{\mathrm{T}}\left(A^{\mathrm{T}}A\otimes BB^{\mathrm{T}}\right)\mathrm{vec}(\Omega-\Omega_m)=\mathrm{tr}\left\{A^{\mathrm{T}}A(\Omega-\Omega_m)BB^{\mathrm{T}}(\Omega-\Omega_m)\right\},$$

we can rewrite (4.8) as

$$\Omega_{m+1} = \underset{\Omega \in \mathbb{S}_p^+}{\arg\min} \left[ \mathcal{F}_\rho(\Omega, \Theta_m, \Gamma_m) + \frac{\rho\tau}{2}\|\Omega - \Omega_m\|_F^2 - \frac{\rho}{2}\mathrm{tr}\left\{A^{\mathrm{T}}A(\Omega - \Omega_m)BB^{\mathrm{T}}(\Omega - \Omega_m)\right\}\right],$$

which is equivalent to

$$\Omega_{m+1} = \underset{\Omega \in \mathbb{S}_p^+}{\arg\min} \left[ \mathrm{tr}\left\{(S + G_m)\,\Omega\right\} - \log\det(\Omega) + \frac{\rho\tau}{2}\|\Omega - \Omega_m\|_F^2 \right], \tag{4.9}$$

where $G_m = \rho A^{\mathrm{T}}(A\Omega_m B - \rho^{-1}\Gamma_m - \Theta_m - C)B^{\mathrm{T}}$. The zero gradient equation for (4.9) is

$$S - \Omega_{m+1}^{-1} + \frac{1}{2}\left(G_m + G_m^{\mathrm{T}}\right) + \rho\tau\left(\Omega_{m+1} - \Omega_m\right) = 0, \tag{4.10}$$

which is a quadratic equation that can be solved in closed form (Witten and Tibshirani, 2009; Price et al., 2015). The solution is

$$\Omega_{m+1} = \frac{1}{2\rho\tau}U\left\{-\Psi + \left(\Psi^2 + 4\rho\tau I_p\right)^{1/2}\right\}U^{\mathrm{T}},$$

where $U\Psi U^{\mathrm{T}}$ is the eigendecomposition of $S + 2^{-1}(G_m + G_m^{\mathrm{T}}) - \rho\tau\Omega_m$. Our majorize-minimize approach is a special case of the *prox-linear* alternating direction method of multiplier algorithm (Chen and Teboulle, 1994; Deng and Yin, 2016).

Conveniently, (4.6) also has a closed form solution:

$$\Theta_{m+1} = \mathrm{soft}\left(A\Omega_{m+1}B - \rho^{-1}\Gamma_m - C, \rho^{-1}\lambda\right),$$

where $\mathrm{soft}(x, \tau) = \max\left(|x| - \tau, 0\right)\mathrm{sign}(x)$. To summarize, we solve (4.2) with the following algorithm.

### 4.3.2 Convergence and implementation

Using the same proof technique as in Deng and Yin (2016), one can show that the iterates from Algorithm 2 converge to their optimal values when a solution to (4.4) exists.

**Algorithm 2** Alternating direction method of multipliers algorithm for (4.2)

Initialize $\Omega_{(0)} \in \mathbb{S}_+^p$, $\Theta_{(0)} \in \mathbb{R}^{a\times b}$, $\rho > 0$, and $\tau$ such that $Q$ is positive definite. Set $m = 0$.
Repeat Step 1 - 6 until convergence:

*Step 1.* Compute $G_m = \rho A^{\mathrm{T}}(A\Omega_m B - \rho^{-1}\Gamma_k - \Theta_m - C)B^{\mathrm{T}}$;
*Step 2.* Decompose $S + 2^{-1}(G_m + G_m^{\mathrm{T}}) - \rho\tau\Omega_m = U\Psi U^{\mathrm{T}}$ where $U$ is orthogonal and $\Psi$ is diagonal;
*Step 3.* Set $\Omega_{m+1} = (2\rho\tau)^{-1}U\left\{-\Psi + (\Psi^2 + 4\rho\tau I_p)^{1/2}\right\}U^{\mathrm{T}}$;
*Step 4.* Set $\Theta_{m+1} = \mathrm{soft}(A\Omega_{m+1}B - \rho^{-1}\Gamma_m - C, \rho^{-1}\lambda)$;
*Step 5.* Set $\Gamma_{m+1} = \Gamma_m - \rho\,(A\Omega_{m+1}B - \Theta_{m+1} - C)$;
*Step 6.* Replace $m$ with $m + 1$.

In our implementation, we set $\tau = \varphi_1(A^{\mathrm{T}}A)\varphi_1(BB^{\mathrm{T}}) + 10^{-8}$, where $\varphi_1(\cdot)$ denotes the largest eigenvalue of its argument. This computation is only needed once at the initialization of our algorithm. We expect that in practice, the computational complexity of our algorithm will be dominated by the eigendecomposition in Step 2, which requires $O(p^3)$ flops.

To select the tuning parameter to use in practice, we recommend using some type of cross-validation procedure based on the application. For example, in the linear discriminant analysis case, one could select the tuning parameter that minimizes the validation misclassification rate or maximizes a validation likelihood.

## 4.4 Statistical Properties

In this section, we show that by using the penalty in (4.2), we can estimate $\Omega_*$ and $A\Omega_*B$ consistently in the Frobenius and $L_1$ norms, respectively. Our results rely on assuming that $A\Omega_*B$ is sparse. Define the set $\mathcal{G}$ as the indices of $A\Omega_*B$ that are nonzero, i.e.,

$$\mathcal{G} = \left\{(i,j) \in \{1,\ldots,a\} \times \{1,\ldots,b\} : [A\Omega_*B]_{ij} \neq 0\right\}.$$

Let the notation $[A\Omega_*B]_{\mathcal{G}} \in \mathbb{R}^{a\times b}$ denote the matrix whose $(i,j)$th entry is equal to the $(i,j)$th of $A\Omega_*B$ if $(i,j) \in \mathcal{G}$ and is equal to zero if $(i,j) \notin \mathcal{G}$. We generalize our results to the case that $A$ and $B$ are unknown, and we use plug-in estimators of them in (4.2).

We first establish convergence rates for the case that $A$ and $B$ are known. Let $\sigma_j(\cdot)$ and

$\varphi_j(\cdot)$ denote the $j$th largest singular value and eigenvalue of their arguments respectively. Suppose that the sample covariance matrix used in (4.2) is $S_n = n^{-1}\sum_{i=1}^{n} X_i X_i^{\mathrm{T}}$, where $X_1, \ldots, X_n$ are independent and identically distributed $p_n$-dimensional random vectors with mean zero and covariance matrix $\Omega_*^{-1}$. We will make the following assumptions:

**Assumption 1**

*For all $n$, there exists a constant $k_1$ such that*

$$0 < k_1^{-1} \leq \varphi_{p_n}(\Omega_*) \leq \varphi_1(\Omega_*) \leq k_1 < \infty.$$

**Assumption 2**

*For all $n$, there exists a constant $k_2$ such that* $\min\{\sigma_{p_n}(A), \sigma_{p_n}(B)\} \geq k_2 > 0$.

**Assumption 3**

*For all $n$, there exist positive constants $k_3$ and $k_4$ such that*

$$\max_{j \in \{1,\ldots,p_n\}} E\left\{\exp(tX_{1j}^2)\right\} \leq k_3 < \infty \text{ for all } t \in (-k_4, k_4).$$

Assumptions 1 and 3 are common in the regularized precision matrix estimation literature, e.g., Assumption 1 was made by Bickel and Levina (2008), Rothman et al. (2008) and Lam and Fan (2009) and Assumption 3 holds if $X_1$ is multivariate normal. Assumption 2 requires that $A$ and $B$ are both rank $p_n$, which has the effect of shrinking every entry of $\hat{\Omega}$. The convergence rate bounds we establish also depend on the quantity

$$\xi(p_n, \mathcal{G}) = \sup_{M \in \mathbb{S}^{p_n}, M \neq 0} \frac{|\,[AMB]_{\mathcal{G}}\,|_1}{\|M\|_F},$$

where $\mathbb{S}^{p_n}$ is the set of symmetric $p_n \times p_n$ matrices. Negahban et al. (2012) defined a similar and more general quantity and called it a compatibility constant.

**Theorem 1**

*Suppose Assumptions 1–3 are true. If $\lambda_n = K_1(n^{-1}\log p_n)^{1/2}$, $K_1$ is sufficiently large, and*

$\xi^2(p_n, \mathcal{G}) \log p_n = o(n)$, *then (i)* $\|\hat{\Omega} - \Omega_*\|_F = O_P\{\xi(p_n, \mathcal{G})(\log p_n/n)^{1/2}\}$ *and*
*(ii)* $|A\hat{\Omega}B - A\Omega_* B|_1 = O_P\{\xi^2(p_n, \mathcal{G})(\log p_n/n)^{1/2}\}$.

The quantity $\xi(p_n, \mathcal{G})$ can be used to recover known results for special cases of (4.2). For example, when $A$ and $B$ are identity matrices, $\xi(p_n, \mathcal{G}) = {s_n}^{1/2}$, where $s_n$ is the number of nonzero entries in $\Omega_*$. This special case was established by Rothman et al. (2008). We can simplify the results of Theorem 1 for case that $A\Omega_* B$ has $g_n$ nonzero entries by introducing an additional assumption:

**Assumption 4**

*For all $n$, there exists a constant $k_5$ such that*

$$\sup_{M \in \mathbb{S}^{p_n}, M \neq 0} \frac{\|[AMB]_{\mathcal{G}}\|_F}{\|M\|_F} \leq k_5 < \infty.$$

Assumption 4 is not the same as bounding $\xi(p_n, \mathcal{G})$ because the numerator uses the Frobenius norm instead of the $L_1$ norm. This requires that for those entries of $A\Omega_* B$ which are nonzero, the corresponding rows and columns of $A$ and $B$, respectively, do not have magnitudes too large as $p_n$ grows.

**Remark 1**

*Assume that the conditions of Theorem 1 are true. If Assumption 4 is true and $A\Omega_* B$ has $g_n$ nonzero entries, then* $\|\hat{\Omega} - \Omega_*\|_F = O_P\{(g_n \log p_n/n)^{1/2}\}$ *and* $|A\hat{\Omega}B - A\Omega_* B|_1 = O_P\{(g_n^2 \log p_n/n)^{1/2}\}$.

In practice, $A$ and $B$ are often unknown and must be estimated. Let $\hat{A}_n$ and $\hat{B}_n$ be estimators of $A$ and $B$. In this case, we estimate $A\Omega_* B$ with $\hat{A}_n \tilde{\Omega} \hat{B}_n$, where

$$\tilde{\Omega} = \underset{\Omega \in \mathbb{S}_+^p}{\arg\min} \left\{ \operatorname{tr}(S_n \Omega) - \log\det(\Omega) + \lambda_n |\hat{A}_n \Omega \hat{B}_n|_1 \right\}. \tag{4.11}$$

Suppose that there exist sequences $a_n$ and $b_n$ such that $|(A - \hat{A}_n)A^+|_1 = O_P(a_n)$ and $|B^+(B - \hat{B}_n)|_1 = O_P(b_n)$, where $A^+$ and $B^+$ are the Moore–Penrose pseudoinverses of $A$ and $B$, respectively; and $a_n = o(1)$, and $b_n = o(1)$.

**Theorem 2**

*Suppose Assumptions 1–3 are true. Let* $C_n = \max\{a_n, b_n\} \,|\, [A\Omega_* B]_{\mathcal{G}} \,|_1$. *If* $\lambda_n = K_2(n^{-1}\log p_n)^{1/2}$, $K_2$ *is sufficiently large,* $\max\{a_n, b_n\} = o(1)$, $\xi^2(p_n, \mathcal{G})\log p_n = o(n)$, *and* $C_n^2 \log p_n = o(n)$, *then (i)* $\|\tilde{\Omega} - \Omega\|_F = O_P\{\xi(p_n, \mathcal{G})(\log p_n/n)^{1/2} + C_n^{1/2}(\log p_n/n)^{1/4}\}$ *and*
*(ii)* $|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_* B|_1 = O_P\{\xi^2(p_n, \mathcal{G})(\log p_n/n)^{1/2} + C_n^{1/2}\xi(p_n, \mathcal{G})(\log p_n/n)^{1/4} + C_n\}$.

The convergence rate bounds in Theorem 2 $(i)$ and $(ii)$ are the sum of the statistical errors from Theorem 1 $(i)$ and $(ii)$ respectively, plus additional error which comes from estimating $A$ and $B$. Proofs of Theorem 1 and Theorem 2 are given in Appendix A.2.

## 4.5 Simulation studies

### 4.5.1 Models

We compare our precision matrix estimator to competing estimators when they are used to fit the linear discriminant analysis model. For 100 independent replications, we generated a realization of $n$ independent copies of $(X, Y)$ defined in (4.3), where $\mu_{*j} = \Omega_*^{-1}\beta_{*j}$ and $P(Y = j) = 1/J$ for $j = 1, \ldots, J$. Using this construction, if the $k$th element of $\beta_{*l} - \beta_{*m}$ is zero, i.e., $\Omega_*(\mu_{*l} - \mu_{*m})$ is zero, then the $k$th variable is non-informative for discriminating between response categories $l$ and $m$.

For each $J \in \{3, \ldots, 10\}$, we partition our $n$ observations into a training set of size $25J$, a validation set of size 200, and a test set of size 1000. We considered two models for $\Omega_*^{-1}$ and $\beta_{*j}$. Let $\mathbb{1}(\cdot)$ be the indicator function.

*Model 1.* We set $\beta_{*j,k} = 1.5\,\mathbb{1}\left[k \in \{4(j-1)+1, \ldots, 4j\}\right]$, so that for any pair of response categories, only eight variables were informative for discrimination. We set $\Omega_{*a,b}^{-1} = .9^{|a-b|}$, so that $\Omega_*$ was tridiagonal.

*Model 2.* We set $\beta_{*j,k} = 2\,\mathbb{1}\left[k \in \{5(j-1)+1, \ldots, 5j\}\right]$, so that for any pair of response categories, only ten variables were informative for discrimination. We set $\Omega_*^{-1}$ to be block diagonal: the block corresponding to the informative variables, i.e., the first $5J$ variables,

had off diagonal entries equal to 0.5 and diagonal entries equal to one. The block submatrix corresponding to the $p-5J$ non-informative variables had $(a,b)$th entry equal to $0.5^{|a-b|}$.

For both models, sparse estimators of $\Omega_*$ should perform well because the precision matrices are very sparse. The total number of informative variables is $4J$ and $5J$ in Models 1 and 2 respectively, so a method like that proposed by Mai et al. (2015), which selects variables that are informative for all pairwise comparisons, may perform poorly when $J$ is large.

### 4.5.2 Methods

We compared several methods in terms of classification accuracy on the test set. We fit (4.3) using the following methods: the sparse naïve Bayes estimator proposed by Guo (2010) with tuning parameter chosen to minimize misclassification rate on the validation set; and the Bayes rule, i.e., $\Omega_*$, $\mu_{*j}$, and $P(Y=j)$ known for $j=1,\ldots,J$. We also fit (4.3) using the ordinary sample means and using the following precision matrix estimators: the estimator we proposed in Section 4.2.2 with tuning parameter chosen to minimize misclassification rate on the validation set; the $L_1$-penalized Gaussian likelihood precision matrix estimator (Yuan and Lin, 2007; Rothman et al., 2008; Friedman et al., 2008) with the tuning parameter chosen to minimize the misclassification rate of the validation set; and a covariance matrix estimator similar to the estimator proposed by Ledoit and Wolf (2004), which is defined by

$$\hat{\Omega}_{\mathrm{LW}}^{-1} = \alpha S + \gamma(1-\alpha)I_p, \tag{4.12}$$

where $(\alpha,\gamma) \in (0,1)\times(0,\infty)$ were chosen to minimize the misclassification rate of the validation set. The $L_1$-penalized Gaussian likelihood precision matrix estimator we used penalized the diagonals. With our data generating models, we found this performed better at classification than (4.1), which does not penalize the diagonals. We also tried two Fisher-linear-discriminant-based methods applicable to multi-category linear discriminant analysis: the sparse linear discriminant method proposed by Witten and Tibshirani (2011)

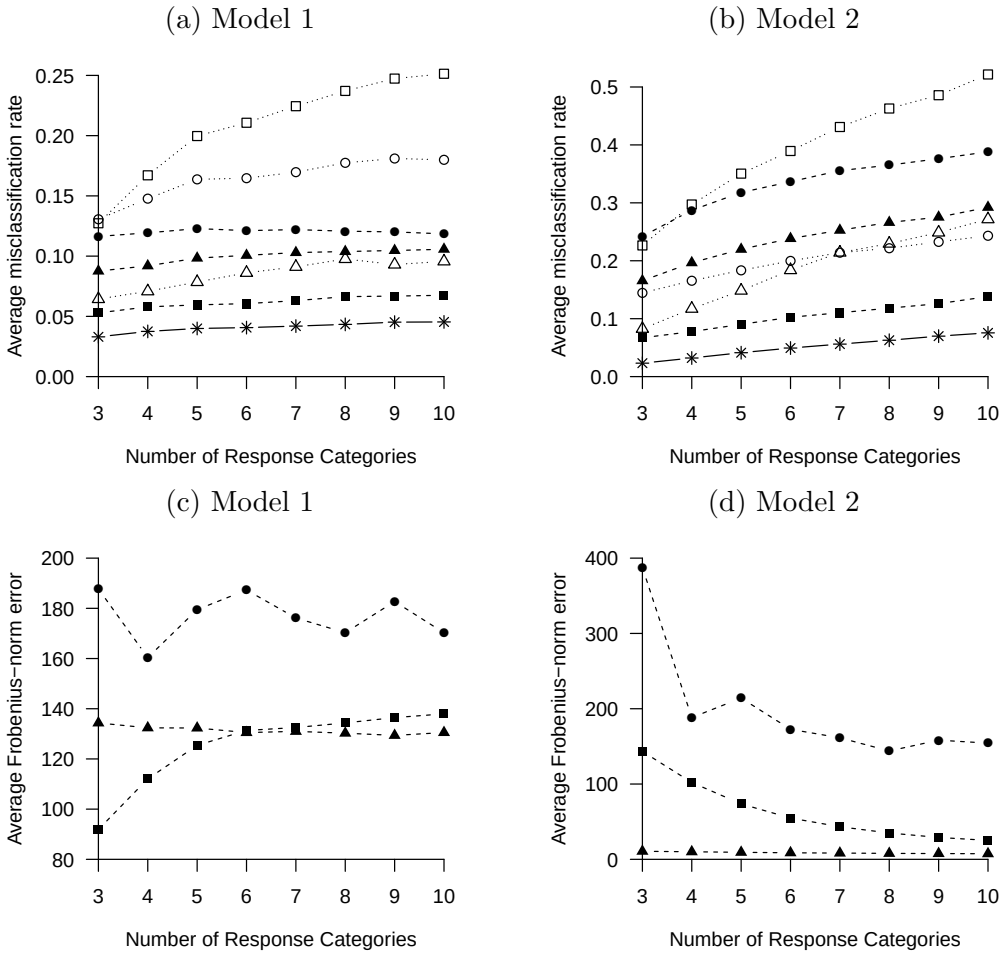



Figure 4.1: Misclassification rates and Frobenius norm error averaged over 100 replications with $p = 200$ for Models 1 and 2. The methods displayed are the estimator we proposed in Section 4.2.2 (dashed and ■), the $L_1$-penalized Gaussian likelihood estimator (dashed and ▲), the Ledoit-Wolf-type estimator from (4.12) (dashed and ●), Bayes (solid and ∗), the method proposed by Guo (2010) (dots and ○), the method proposed by Mai et al. (2015) (dots and △), and the method proposed by Witten and Tibshirani (2011) (dots and □).

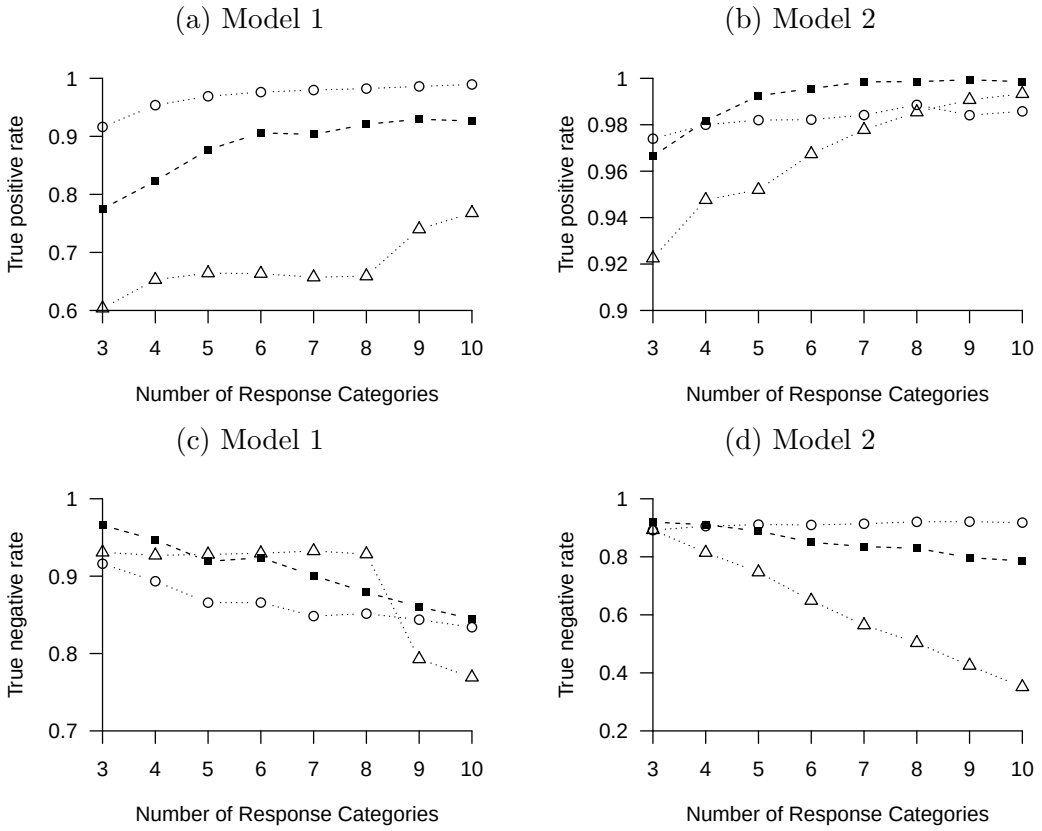



Figure 4.2: True positive and true negative rates averaged over 100 replications with $p = 200$ for Model 1 in (a) and (c); and for Model 2 in (b) and (d). The methods displayed are the estimator we proposed in Section 4.2.2 (dashed and ■), the method proposed by Guo (2010) (dots and ○), and the method proposed by Mai et al. (2015) (dots and △).

with tuning parameter and dimension chosen to minimize the misclassification rate of the validation set; and the multi-category sparse linear discriminant method proposed by Mai et al. (2015) with tuning parameter chosen to minimize the misclassification rate of the validation set.

We could also have selected tuning parameters for the model-based methods by maximizing a validation likelihood or using an information criterion, but minimizing the misclassification rate on a validation set made it fairer to compare the model-based methods and the Fisher-linear-discriminant-based methods in terms of classification accuracy.

### 4.5.3 Performance measures

We measured classification accuracy on the test set for each replication for the methods described in Section 4.5.2. For the methods that produced a precision matrix estimator, we also measured this estimator's Frobenius norm error: $\|\bar{\Omega} - \Omega_*\|_F$, where $\bar{\Omega}$ is the estimator. To measure variable selection accuracy, we used both the true positive rate and the true negative rate, which are respectively defined by

$$\frac{\text{card}\left\{(m,k) : \hat{\Delta}_{m,k} \neq 0 \cap \Delta_{*m,k} \neq 0\right\}}{\text{card}\left\{(m,k) : \Delta_{*m,k} \neq 0\right\}}, \quad \frac{\text{card}\left\{(m,k) : \hat{\Delta}_{m,k} = 0 \cap \Delta_{*m,k} = 0\right\}}{\text{card}\left\{(m,k) : \Delta_{*m,k} = 0\right\}},$$

where $(m,k) \in \{2,\ldots,J\} \times \{1,\ldots,p\}$, $\Delta_{*m} = \beta_{*1} - \beta_{*m}$, $\hat{\Delta}_m$ is an estimator of $\Delta_{*m}$, and card denotes the cardinality of a set.

### 4.5.4 Results

We display average misclassification rates and Frobenius norm error averages for both models with $p = 200$ in Figure 4.1, and display variable selection accuracy averages in Figure 4.2. For both models, our method outperformed all competitors in terms of classification accuracy for all $J$, except the Bayes rule, which uses population parameter values unknown in practice. In terms of precision matrix estimation, for Model 1, our method did better than the $L_1$-penalized Gaussian likelihood precision matrix estimator when the sample size was small, but became worse than the $L_1$-penalized Gaussian likelihood precision matrix estimator as the sample size increased. For Model 2, our method was worse than the $L_1$-penalized Gaussian likelihood precision matrix estimator in Frobenius norm error for precision matrix estimation, but was better in terms of classification accuracy.

In terms of variable selection, our method was competitive with the methods proposed by Guo (2010) and Mai et al. (2015). For Model 1, our method tended to have a higher average true negative rate than the method of Guo (2010) and a lower average true positive rate than the method of Mai et al. (2015). For Model 2, all methods tended to have relatively high average true positive rates, while our method had a higher average true negative rate

than the method of Mai et al. (2015). Although the method proposed by Guo (2010) had a higher average true negative rate for Model 2 than our proposed method had, our method performed better in terms of classification accuracy.

## 4.6 Genomic data example

We used our method to fit the linear discriminant analysis model in a real data application. The data are gene expression profiles consisting of $p = 22,283$ genes from 127 subjects, who either have Crohn's disease, ulcerative colitis, or neither. This dataset comes from Burczynski et al. (2006). The goal of our analysis was to fit a linear discriminant analysis model that could be used to identify which genes are informative for discriminating between each pair of the response categories. These data were also analyzed in Mai et al. (2015). One difference between our method and the method of Mai et al. (2015) is that the method of Mai et al. (2015) excludes variables from all pairwise response category comparisons whereas our method allows a distinct set of informative variables to be estimated for each comparison.

To measure the classification accuracy of our method and its competitors, we randomly split the data into training set of size 100 and test set of size 27 for 100 independent replications. Within each replication, we first applied a screening rule to the training set as in Rothman et al. (2009) and Mai et al. (2015) based on $F$-test statistics, and then restricted our discriminant analysis model to the genes with the $k$ largest $F$-test statistic values.

We chose tuning parameters with 5-fold cross validation that minimized the validation classification error rate. Misclassification rates are shown in Figure 4.3, where we compared our method to the method proposed by Mai et al. (2015), the method proposed by Witten and Tibshirani (2011), the method proposed by Guo (2010), and the method that used the $L_1$-penalized Gaussian likelihood precision matrix estimator. We saw that our method was as or more accurate in terms of classification accuracy than the competing methods. The only method that performed nearly as well was that of Mai et al. (2015) when we

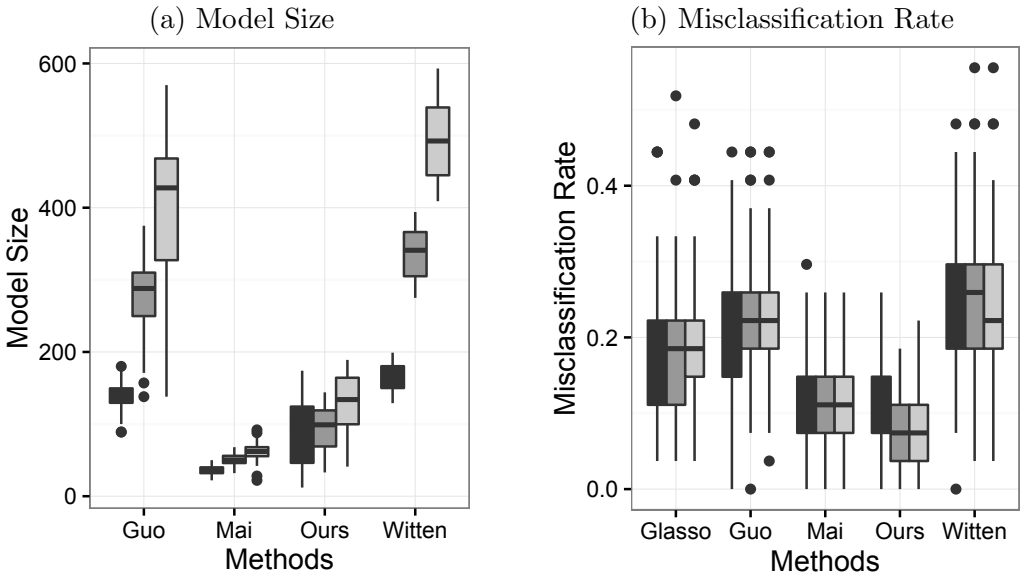



Figure 4.3: Model sizes and misclassification rates from 100 random training/testing splits with $k = 100$ (dark grey), $k = 200$ (grey), and $k = 300$ (light grey). Guo is the method proposed by Guo (2010), Mai is the method proposed by Mai et al. (2015), Glasso is the $L_1$-penalized Gaussian likelihood precision matrix estimator, Ours is the estimator we propose Section 4.2.2, and Witten is the method proposed by Witten and Tibshirani (2011).

used $k = 100$ screened genes. However, the best out-of-sample classification accuracy was achieved with $k = 300$, where our method was significantly better than the competitors.

In Figure 4.3, we also display model sizes, i.e., the total number of variables that were estimated to be informative for discriminating between response categories. To measure model size for the method proposed by Witten and Tibshirani (2011), we used the version of their method with two discriminant vectors. We saw that although the method of Mai et al. (2015) tended to estimate slightly smaller models, our method, which performs best in classification, selects only slightly more variables. Moreover, our method can be used to identify a distinct subset of genes that are informative specifically for discriminating between patients with Crohn's disease and ulcerative colitis. This was of interest in the study of Burczynski et al. (2006).

# References


Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405.

Allen, G. I. and Tibshirani, R. J. (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research*, 9:485–516.

Bhadra, A. and Mallick, B. K. (2013). Joint high-dimensional bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69(2):447–457.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B*, 59(1):3–54.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. (2009). Sparse and stable

Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272.

Burczynski, M. E., Peterson, R. L., Twine, N. C., Zuberek, K. A., Brodeur, B. J., Casciotti, L., Maganti, V., Reddy, P. S., Strahs, A., Immermann, F., et al. (2006). Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *The Journal of Molecular Diagnostics*, 8(1):51–61.

Cai, T. T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144.

Chen, G. and Teboulle, M. (1994). A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64:81–101.

Chen, K., Dong, H., and Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100:902–920.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Chen, X., Xu, M., and Wu, W. B. (2016). Regularized estimation of linear functionals of precision matrices for high-dimensional time series. *IEEE Transactions on Signal Processing*, 64(24):6459–6470.

Chi, E. C. and Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., and Ryder, T. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541.

Cook, R. D., Forzani, L., and Rothman, A. J. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *The Annals of Statistics*, 40(1):353–384.

Cook, R. D., Forzani, L., and Rothman, A. J. (2013). Prediction in abundant high-dimensional linear regression. *Electronic Journal of Statistics*, 7:3059–3088.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica*, 20(3):927–1010.

Dalal, O. and Rajaratnam, B. (2014). G-AMA: Sparse Gaussian graphical model estimation via alternating minimization. *arXiv preprint arXiv:1405.3034*.

Deng, W. and Yin, W. (2016). On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123.

Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B*, 74(4):745–771.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541.

Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32.

Friedman, J. H., Hastie, T. J., and Tibshirani, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441.

Geng, H., Iqbal, J., Chan, W. C., and Ali, H. H. (2011). Virtual CGH: an integrative approach to predict genetic abnormalities from gene expression microarray data applied in lymphoma. *BMC Medical Genomics*, 4(32):1–14.

Goldstein, T., O'Donoghue, B., Setzer, S., and Baraniuk, R. (2014). Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623.

Guo, J. (2010). Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis. *Biostatistics*, 11:599–608.

Gupta, A. K. and Nagar, D. K. (2000). *Matrix variate distributions*. Chapman and Hall/CRC Press, Boca Raton, FL.

Helwig, N. E. (2015). *eegkit: Toolkit for Electroencephalography Data*. R package version 1.0-2.

Hoff, P. D. (2011). Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6(2):179–196.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. K. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*, volume 24, pages 2330–2338. MIT Press, Cambridge, MA.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Hung, H. and Wang, C.-C. (2013). Matrix variate logistic regression model with application to EEG data. *Biostatistics*, 14:189–202.

Hunter, D. R. and Li, R. (2005). Variable selection using mm algorithms. *The Annals of Statistics*, 33(4):1617.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254—4278.

Lange, K. (2016). *MM Optimization Algorithms*. SIAM, Philadelphia, PA.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Lee, W. and Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *Journal of Multivariate Analysis*, 111:241–255.

Leng, C. and Tang, C. Y. (2012). Sparse matrix graphical models. *Journal of the American Statistical Association*, 107(499):1187–1200.

Li, B., Kim, M. K., and Altman, N. (2010). On dimension folding of matrix-or array-valued statistical objects. *The Annals of Statistics*, 38(2):1094–1121.

Li, M. and Yuan, B. (2005). 2D-LDA: A statistical linear discriminant analysis for image matrix. *Pattern Recognition Letters*, 26(5):527–532.

Mai, Q., Yang, Y., and Zou, H. (2015). Multiclass sparse discriminant analysis. *arXiv preprint arXiv:1504.05845*.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99:29–42.

Manceur, A. M. and Dutilleul, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239:37–49.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Mitchell, M. W., Genton, M. G., and Gumpertz, M. L. (2006). A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis*, 97:1025–1043.

Molstad, A. J. and Rothman, A. J. (2016a). Indirect multivariate response linear regression. *Biometrika*, 103:595–607.

Molstad, A. J. and Rothman, A. J. (2016b). A penalized likelihood method for classification with matrix-valued predictors. *arXiv preprint arXiv:1609.07386*.

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons.

Mukherjee, A. and Zhu, J. (2011). Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining*, 4(6):612–622.

Negahban, S. N., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2012). A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Obozinski, G., Taskar, B., and Jordan, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39:1–47.

O'Donoghue, B. and Candes, E. (2015). Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732.

Pan, R., Wang, H., and Li, R. (2016). Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening. *Journal of the American Statistical Association*, 111(513):169–179.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 4(1):53–77.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation: With High-Dimensional Data*. John Wiley & Sons.

Price, B. S., Geyer, C. J., and Rothman, A. J. (2015). Ridge fusion in statistical learning. *Journal of Computational and Graphical Statistics*, 24(2):439–454.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-rank Regression*. Springer, New York, NY.

Roś, B., Bijma, F., de Munck, J. C., and de Gunst, M. C. (2016). Existence and uniqueness of the maximum likelihood estimator for models with a Kronecker product covariance structure. *Journal of Multivariate Analysis*, 143:345–361.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, 99:733–740.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rothman, A. J. and Forzani, L. (2014). On the existence of the weighted bridge penalized Gaussian likelihood precision matrix estimator. *Electronic Journal of Statistics*, 8:2693–2700.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, 98:133–146.

Su, Z. and Cook, R. D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika*, 99:687–702.

Su, Z. and Cook, R. D. (2013). Scaled envelopes: Scale invariant and efficient estimation in multivariate linear regression. *Biometrika*, 100:921–938.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288.

Tseng, P. (1991). Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization*, 29(1):119–138.

Tsiligkaridis, T., Hero, A. O., and Zhou, S. (2012). Kronecker graphical lasso. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 884–887. IEEE.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, 47(3):349–363.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Witten, D. M. and Tibshirani, R. J. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B*, 71(3):615–636.

Witten, D. M. and Tibshirani, R. J. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B*, 73(5):753–772.

Witten, D. M., Tibshirani, R. J., and Hastie, T. J. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Xu, P., Zhu, J., Zhu, L., and Li, Y. (2015). Covariance-enhanced discriminant analysis. *Biometrika*, 102:33–45.

Xue, L., Ma, S., and Zou, H. (2012). Positive-definite $\ell_1$-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500):1480–1491.

Yuan, M. (2008). Efficient computation of $\ell_1$ regularized estimates in gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):809–826.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B*, 69(3):329–346.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 93:19–35.

Yuan, Y., Curtis, C., Caldas, C., and Markowetz, F. (2012). A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):947–954.

Zhang, Y. and Schneider, J. G. (2010). Learning multiple tasks with a sparse matrix-normal penalty. In *Advances in Neural Information Processing Systems*, pages 2550–2558.

Zhong, W. and Suslick, K. S. (2015). Matrix discriminant analysis with application to colorimetric sensor array data. *Technometrics*, 57(4):524–534.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B*, 76(2):463–483.

Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562.

# Appendix A

# Proofs

## A.1 Proofs for Chapter 3

**Proof of Proposition 1.** Since $\Sigma_*$ is positive definite, we apply the partitioned inverse formula to obtain

$$\Sigma_*^{-1} = \begin{pmatrix} \Sigma_{*XX} & \Sigma_{*XY} \\ \Sigma_{*XY}^{\mathrm{T}} & \Sigma_{*YY} \end{pmatrix}^{-1} = \begin{pmatrix} \Delta_*^{-1} & -\beta_* \Sigma_{*E}^{-1} \\ -\eta_* \Delta_*^{-1} & \Sigma_{*E}^{-1} \end{pmatrix},$$

where $\Delta_* = \Sigma_{*XX} - \Sigma_{*XY}\Sigma_{*YY}^{-1}\Sigma_{*XY}^{\mathrm{T}}$ and $\Sigma_{*E} = \Sigma_{*YY} - \Sigma_{*XY}^{\mathrm{T}}\Sigma_{*XX}^{-1}\Sigma_{*XY}$. The symmetry of $\Sigma_*^{-1}$ implies that $\beta_*\Sigma_{*E}^{-1} = (\eta_*\Delta_*^{-1})^{\mathrm{T}}$ so

$$\beta_* = \Delta_*^{-1}\eta_*^{\mathrm{T}}\Sigma_{*E}. \tag{A.1}$$

Using the Woodbury identity,

$$\begin{aligned} \Sigma_{*E}^{-1} &= (\Sigma_{*YY} - \Sigma_{*XY}^{\mathrm{T}}\Sigma_{*XX}^{-1}\Sigma_{*XY})^{-1} \\ &= \Sigma_{*YY}^{-1} + \Sigma_{*YY}^{-1}\Sigma_{*XY}^{\mathrm{T}}\left(\Sigma_{*XX}^{-1} - \Sigma_{*XY}\Sigma_{*YY}^{-1}\Sigma_{*XY}^{\mathrm{T}}\right)^{-1}\Sigma_{XY}\Sigma_{*YY}^{-1} \\ &= \Sigma_{*YY}^{-1} + \eta_*\Delta_*^{-1}\eta_*^{\mathrm{T}}. \end{aligned} \tag{A.2}$$

Using the inverse of the expression above in (A.1) establishes the result. □

In our proof of Proposition 2, we use the matrix inequality

$$\begin{aligned}\|A^{(1)}A^{(2)}A^{(3)} - B^{(1)}B^{(2)}B^{(3)}\| \leq \sum_{j=1}^{3} \|A^{(j)} - B^{(j)}\| \prod_{k\neq j} \|B^{(k)}\| \\ + \sum_{j=1}^{3} \|B^{(j)}\| \prod_{k\neq j} \|A^{(k)} - B^{(k)}\| + \prod_{j=1}^{3} \|A^{(j)} - B^{(j)}\|.\end{aligned} \tag{A.3}$$

Bickel and Levina (2008) used (A.3) to prove their Theorem 3.

**Proof of Proposition 2.** From (A.2) in the proof of Proposition 1, $\Sigma_{*E}^{-1} = \Sigma_{*YY}^{-1} + \eta_*\Delta_*^{-1}\eta_*^{\mathrm{T}}$. Define $\hat{\Sigma}_E^{-1} = \hat{\Sigma}_{YY}^{-1} + \hat{\eta}\hat{\Delta}^{-1}\hat{\eta}^{\mathrm{T}}$. Applying (A.3),

$$\begin{aligned}\|\hat{\beta} - \beta_*\| =& \|\hat{\Delta}^{-1}\hat{\eta}^{\mathrm{T}}\hat{\Sigma}_E - \Delta_*^{-1}\eta_*^{\mathrm{T}}\Sigma_{*E}\| \\ \leq& \|\hat{\Delta}^{-1} - \Delta_*^{-1}\|\|\eta_*\|\|\Sigma_{*E}\| + \|\hat{\eta} - \eta_*\|\|\Delta_*^{-1}\|\|\Sigma_{*E}\| + \|\hat{\Sigma}_E - \Sigma_{*E}\|\|\Delta_*^{-1}\|\|\eta_*\| \\ &+ \|\Delta_*^{-1}\|\|\hat{\eta} - \eta_*\|\|\hat{\Sigma}_E - \Sigma_{*E}\| + \|\eta_*\|\|\hat{\Delta}^{-1} - \Delta_*^{-1}\|\|\hat{\Sigma}_E - \Sigma_{*E}\| \\ &+ \|\Sigma_{*E}\|\|\hat{\Delta}^{-1} - \Delta_*^{-1}\|\|\hat{\eta} - \eta_*\| + \|\hat{\eta} - \eta_*\|\|\hat{\Delta}^{-1} - \Delta_*^{-1}\|\|\hat{\Sigma}_E - \Sigma_{*E}\|.\end{aligned} \tag{A.4}$$

We will show that the third term in (A.4) dominates the others. We continue by deriving its bound. Employing a matrix identity used by Cai et al. (2010), we write $\hat{\Sigma}_E - \Sigma_{*E} = \Sigma_{*E}(\Sigma_{*E}^{-1} - \hat{\Sigma}_E^{-1})\hat{\Sigma}_E$, so

$$\|\hat{\Sigma}_E - \Sigma_{*E}\| \leq \|\hat{\Sigma}_E\|\|\Sigma_{*E}\|\|\hat{\Sigma}_E^{-1} - \Sigma_{*E}^{-1}\|. \tag{A.5}$$

Using the triangle inequality and (A.3),

$$\begin{aligned}\|\hat{\Sigma}_E^{-1} - \Sigma_{*E}^{-1}\| &\leq \|\hat{\Sigma}_{YY}^{-1} - \Sigma_{*YY}^{-1}\| + \|\hat{\eta}\hat{\Delta}^{-1}\hat{\eta}^{\mathrm{T}} - \eta_*\Delta_*^{-1}\eta_*^{\mathrm{T}}\| \\ &\leq \|\hat{\Sigma}_{YY}^{-1} - \Sigma_{*YY}^{-1}\| + 2\|\hat{\eta} - \eta_*\|\|\Delta_*^{-1}\|\|\eta_*\| + \|\hat{\Delta}^{-1} - \Delta_*^{-1}\|\|\eta_*\|^2 \\ &\quad + 2\|\eta_*\|\|\hat{\Delta}^{-1} - \Delta_*^{-1}\|\|\hat{\eta} - \eta_*\| + \|\Delta_*^{-1}\|\|\hat{\eta} - \eta_*\|^2 + \|\hat{\eta} - \eta_*\|^2\|\hat{\Delta}^{-1} - \Delta_*^{-1}\| \\ &= O_P\left(c_n + a_n\|\eta_*\|\|\Delta_*^{-1}\| + b_n\|\eta_*\|^2\right).\end{aligned} \tag{A.6}$$

Since $\varphi_{\min}(\Sigma_{*YY}^{-1}) \geq K$ and $\Delta_*^{-1}$ is positive definite, Weyl's eigenvalue inequality implies

that $\varphi_{\min}(\Sigma_{*E}^{-1}) \geq K$ so

$$\|\Sigma_{*E}\| = \varphi_{\min}^{-1}(\Sigma_{*E}^{-1}) \leq 1/K. \tag{A.7}$$

Also,

$$\|\hat{\Sigma}_E\| = \varphi_{\min}^{-1}(\hat{\Sigma}_E^{-1}) = O_P(1) \tag{A.8}$$

because $\varphi_{\min}(\Sigma_{*E}^{-1}) \geq K$, $\hat{\Sigma}_E$ is positive definite, and $a_n\|\eta_*\|\|\Delta_*^{-1}\| + b_n\|\eta_*\|^2 + c_n = o(1)$ in (A.6). Using (A.6), (A.7), and (A.8), in (A.5),

$$\|\hat{\Sigma}_E - \Sigma_{*E}\| = O_P\left(a_n\|\eta_*\|\|\Delta_*^{-1}\| + b_n\|\eta_*\|^2 + c_n\right).$$

We then see that the third term in (A.4) dominates and

$$\begin{aligned}\|\hat{\beta} - \beta_*\| &= O_P\left\{\left(a_n\|\eta_*\|\|\Delta_*^{-1}\| + b_n\|\eta_*\|^2 + c_n\right)\|\eta_*\|\|\Delta_*^{-1}\|\right\} \\ &= O_P\left(a_n\|\eta_*\|^2\|\Delta_*^{-1}\|^2 + b_n\|\eta_*\|^3\|\Delta_*^{-1}\| + c_n\|\eta_*\|\|\Delta_*^{-1}\|\right).\end{aligned}$$

## A.2 Proofs for Chapter 4

### A.2.1 Notation

Define the following norms: $\|A\|_\infty = \max_{i,j}|A_{ij}|$, $|A|_1 = \sum_{i,j}|A_{ij}|$, $\|A\|_F = \mathrm{tr}(A^{\mathrm{T}}A)$, $\|A\| = \sigma_1(A)$. Let $\mathbb{S}^p$ denote the set of $p \times p$ symmetric matrices. To simplify notation, we omit the subscript $n$ from $S_n$, $\lambda_n$, and $p_n$ as defined in Section 4.4 and let $\kappa = k_1^{-2}$.

### A.2.2 Proof of Theorem 1

To prove Theorem 1, we use a strategy similar to that employed by Rothman (2012).

**Lemma 1**
*Suppose that Assumptions 1–3 hold, and* $\lambda \leq \epsilon\kappa\{\xi(p,\mathcal{G})\tau\}^{-1}$ *for some* $\tau > 12$. *Then for all positive and sufficiently small* $\epsilon$, $\|B^+(S - \Omega_*^{-1})A^+\|_\infty \leq \lambda/2$ *implies* $\|\hat{\Omega} - \Omega_*\|_F \leq \epsilon$.

**Proof of Lemma 1.** We follow the proof techniques used by Rothman et al. (2008), Negahban et al. (2012) and Rothman (2012). Define $B_\epsilon = \{\Delta \in \mathbb{S}^p : \|\Delta\|_F \leq \epsilon\}$. Let $f$ be the objective function in (4.2). Because $f$ is convex and $\hat{\Omega}$ is its minimizer, $\inf\{f(\Omega_* + \Delta) : \Delta \in B_\epsilon\} >$

$f(\Omega_*)$, implies $\|\hat{\Omega} - \Omega_*\|_F \leq \epsilon$ (Rothman et al., 2008). Define $D(\Delta) = f(\Omega_* + \Delta) - f(\Omega_*)$. Then

$$D(\Delta) = \text{tr}(S\Delta) + \log\det(\Omega_*) - \log\det(\Omega_* + \Delta) + \lambda_1 \left\{|A(\Omega_* + \Delta)B|_1 - |A\Omega_* B|_1\right\}.$$

By the arguments used in Rothman et al. (2008), $\log\det(\Omega_*) - \log\det(\Omega_* + \Delta) \geq -\text{tr}(S\Omega_*^{-1}) + 8^{-1}\kappa\|\Delta\|_F^2$, so that

$$D(\Delta) \geq \text{tr}\left\{\Delta(S - \Omega_*^{-1})\right\} + \frac{1}{8}\kappa\|\Delta\|_F^2 + \lambda_1 \left\{|A(\Omega_* + \Delta)B|_1 - |A\Omega_* B|_1\right\}. \tag{A.9}$$

We now bound $|A(\Omega_* + \Delta)B|_1 - |A\Omega_* B|_1$ in (A.9). Recall that

$$\mathcal{G} = \{(i,j) \in \{1,\ldots,a\} \times \{1,\ldots,b\} : [A\Omega_* B]_{ij} \neq 0\}$$

and $\mathcal{G}^c = \{1,\ldots,a\} \times \{1,\ldots,b\} \setminus \mathcal{G}$. Since $|A\Omega_* B|_1 = |\,[A\Omega_* B]_{\mathcal{G}}\,|_1$ and $|A(\Omega_* + \Delta)B|_1 = |\,[A\Omega_* B]_{\mathcal{G}} + [A\Delta B]_{\mathcal{G}}\,|_1 + |\,[A\Delta B]_{\mathcal{G}^c}\,|_1$, we can apply the reverse triangle inequality: $|A(\Omega_* + \Delta)B|_1 - |A\Omega_* B|_1 \geq |\,[A\Delta B]_{\mathcal{G}^c}\,|_1 - |\,[A\Delta B]_{\mathcal{G}}\,|_1$. Plugging this bound into (A.9),

$$D(\Delta) \geq \text{tr}\left\{(S - \Omega_*^{-1})\Delta\right\} + \frac{1}{8}\kappa\|\Delta\|_F^2 + \lambda_1 \left(|\,[A\Delta B]_{\mathcal{G}^c}\,|_1 - |\,[A\Delta B]_{\mathcal{G}}\,|_1\right). \tag{A.10}$$

We now bound $\text{tr}\left\{(S - \Omega_*^{-1})\Delta\right\}$. Let $A^+ = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ and $B^+ = B^{\mathrm{T}}(BB^{\mathrm{T}})^{-1}$. Because $A$ and $B$ are both rank $p$ by Assumption 2, $A^+A = I_p$ and $BB^+ = I_p$. Thus

$$\begin{aligned}
\text{tr}\left\{(S - \Omega_*^{-1})\Delta\right\} &\geq -|\text{tr}\left\{(S - \Omega_*^{-1})\Delta\right\}| = -|\text{tr}\left\{(S - \Omega_*^{-1})A^+A\Delta BB^+\right\}| \\
&= -|\text{tr}\left\{B^+(S - \Omega_*^{-1})A^+A\Delta B\right\}| \\
&\geq -\|B^+(S - \Omega_*^{-1})A^+\|_\infty |A\Delta B|_1. \qquad \text{(A.11)}
\end{aligned}$$

By assumption, $\|B^+(S - \Omega_*^{-1})A^+\|_\infty \leq \lambda/2$, so applying (A.11) to (A.10),

$$\begin{aligned}
D(\Delta) &\geq \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{\lambda}{2}|A\Delta B|_1 + \lambda\left(|\,[A\Delta B]_{\mathcal{G}^c}\,|_1 - |\,[A\Delta B]_{\mathcal{G}}\,|_1\right) \qquad \text{(A.12)} \\
&= \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{\lambda}{2}\left(|\,[A\Delta B]_{\mathcal{G}}\,|_1 + |\,[A\Delta B]_{\mathcal{G}^c}\,|_1\right) + \lambda\left(|\,[A\Delta B]_{\mathcal{G}^c}\,|_1 - |\,[A\Delta B]_{\mathcal{G}}\,|_1\right) \\
&= \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{3}{2}\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1 + \frac{1}{2}\lambda|\,[A\Delta B]_{\mathcal{G}^c}\,|_1 \\
&\geq \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{3}{2}\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1. \qquad \text{(A.13)}
\end{aligned}$$

We now bound the quantity $|\,[A\Delta B]_{\mathcal{G}}\,|_1$. Multiplying and dividing $\Delta$ by $\|\Delta\|_F$,

$$\left|\left[A\frac{\|\Delta\|_F}{\|\Delta\|_F}\Delta B\right]_{\mathcal{G}}\right|_1 = \|\Delta\|_F\left|\left[A\frac{1}{\|\Delta\|_F}\Delta B\right]_{\mathcal{G}}\right|_1 \leq \|\Delta\|_F\left(\sup_{M\in\mathbb{S}^p, M\neq 0}\frac{|\,[AMB]_{\mathcal{G}}\,|_1}{\|M\|_F}\right),$$

so that $|\,[A\Delta B]_{\mathcal{G}}\,|_1 \leq \|\Delta\|_F\,\xi(p,\mathcal{G})$. Finally, since $\lambda \leq \epsilon\kappa\,\{\tau\xi(p,\mathcal{G})\}^{-1}$ with $\tau > 12$, $\|\Delta\|_F = \epsilon$ for $\Delta \in \mathcal{B}_\epsilon$,

$$\begin{aligned} D(\Delta) &\geq \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{3}{2}\lambda\|\Delta\|_F\xi(p,\mathcal{G}) \\ &= \|\Delta\|_F^2\left\{\frac{1}{8}\kappa - \frac{3\lambda\xi(p,\mathcal{G})}{2\|\Delta\|_F}\right\} \geq \epsilon^2\left(\frac{1}{12}\kappa - \frac{1}{\tau}\kappa\right) > 0. \end{aligned}$$

which establishes the desired result.

The following lemma follows from the proof of Lemma 1 of Negahban et al. (2012).

**Lemma 2**

*If the conditions of Lemma 1 are true, then $\hat{\Delta} = \hat{\Omega} - \Omega_*$ belongs to the set*

$$\left\{\Delta \in \mathbb{S}^p : |\,[A\Delta B]_{\mathcal{G}^c}\,|_1 \leq 3|\,[A\Delta B]_{\mathcal{G}}\,|_1\right\}.$$

Lemma 3 follows from the proof of Lemma 2 from Lam and Fan (2009), Assumption 2, and Lemma A.3 of Bickel and Levina (2008).

**Lemma 3**

*Suppose Assumptions 1–3 hold. Then, there exist constants $C_1$ and $C_2$ such that*

$$P(\|B^+SA^+ - B^+\Omega_*^{-1}A^+\|_\infty \geq \nu) \leq C_1p^2\exp(-C_2n\nu^2),$$

*for $|\nu| \leq \delta$ where $C_1, C_2$, and $\delta$ do not depend on $n$.*

**Proof of Theorem 1.** Set $\epsilon = K_1\kappa^{-1}\xi(p,\mathcal{G})(n^{-1}\log p)^{1/2}$, $\lambda = K_1\tau_1^{-1}(n^{-1}\log p)^{1/2}$ with $\tau_1 > 12$. Applying Lemma 1 and Lemma 3, there exist constants $C_1$ and $C_2$ such that for

sufficiently large $n$,

$$P\left(\|\hat{\Omega}-\Omega_*\|_F \leq K_1\kappa^{-1}\xi(p,\mathcal{G})\sqrt{\frac{\log p}{n}}\right) \geq P\left(\|B^+(S-\Omega_*^{-1})A^+\|_\infty \leq \frac{K_1}{2\tau_1}\sqrt{\frac{\log p}{n}}\right)$$
$$\geq 1 - C_1 p^{2-C_2K_1/2\tau_1},$$

which establishes $(i)$ because $1 - C_1 p^{2-C_2K_1/2\tau_1} \to 1$ as $K_1 \to \infty$. To establish $(ii)$,

$$\begin{aligned}|A(\hat{\Omega}-\Omega_*)B|_1 &= |[A(\hat{\Omega}-\Omega_*)B]_{\mathcal{G}}|_1 + |[A(\hat{\Omega}-\Omega_*)B]_{\mathcal{G}^c}|_1 \\ &\leq 4|[A(\hat{\Omega}-\Omega_*)B]_{\mathcal{G}}|_1 \qquad \text{(A.14)} \\ &\leq 4\xi(p,\mathcal{G})\|\hat{\Omega}-\Omega_*\|_F, \qquad \text{(A.15)}\end{aligned}$$

where (A.14) follows from Lemma 2 and (A.15) follows from the definition of $\xi(p,\mathcal{G})$.

### A.2.3 Proof of Theorem 2

**Lemma 4**

*Let $C_a$ and $C_b$ be constants. Let $a_n$ and $b_n$ be sequences such that $|(\hat{A}_n - A)A^+|_1 \leq C_a a_n$ and $|B^+(\hat{B}_n - B)|_1 \leq C_b b_n$ with probability at least $1 - f(C_a)$ and $1 - g(C_b)$. Let $d_n = C_a a_n + C_b b_n + C_a C_b a_n b_n$. Then*

$$|\hat{A}_n(\Omega_* + \Delta)\hat{B}_n|_1 - |\hat{A}_n\Omega_*\hat{B}_n|_1 \geq |A(\Delta+\Omega_*)B|_1 - |[A\Omega_*B]_{\mathcal{G}}|_1$$
$$+ d_n\left(|A\Delta B|_1 + 2|[A\Omega_*B]_{\mathcal{G}}|_1\right),$$

*with probability at least* $\min\{1 - f(C_a), 1 - g(C_b)\}$.

**Proof of Lemma 4.** Let $|\hat{A}_n(\Omega_* + \Delta)\hat{B}_n|_1 - |\hat{A}_n\Omega_*\hat{B}_n|_1 \equiv V_1 - V_2$. First,

$$\begin{aligned}V_1 &= |\hat{A}_n(\Omega_*+\Delta)\hat{B}_n + A(\Omega_*+\Delta)B - A(\Omega_*+\Delta)B|_1 \\ &\geq |A(\Omega_*+\Delta)B|_1 - |A(\Omega_*+\Delta)B - \hat{A}_n(\Omega_*+\Delta)\hat{B}_n|_1, \qquad \text{(A.16)}\end{aligned}$$

by the triangle inequality. Also,

$$V_2 = |\hat{A}_n\Omega_*\hat{B}_n - A\Omega_*B + A\Omega_*B|_1 \leq |\hat{A}_n\Omega_*\hat{B}_n - [A\Omega_*B]_{\mathcal{G}}|_1 + |[A\Omega_*B]_{\mathcal{G}}|_1, \qquad \text{(A.17)}$$

so that from (A.16) and (A.17),

$$\begin{aligned} V_1 - V_2 \geq & |A(\Omega_* + \Delta)B|_1 - |[A\Omega_* B]_{\mathcal{G}}|_1 \\ & - |A(\Omega_* + \Delta)B - \hat{A}_n(\Omega_* + \Delta)\hat{B}_n|_1 - |\hat{A}_n\Omega_*\hat{B}_n - A\Omega_* B|_1. \end{aligned} \tag{A.18}$$

Let $V_3 = -|A(\Omega_* + \Delta)B - \hat{A}_n(\Omega_* + \Delta)\hat{B}_n|_1 - |\hat{A}_n\Omega_*\hat{B}_n - A\Omega_* B|_1$. By a triangle inequality on the first term of $V_3$,

$$V_3 \geq -2|\hat{A}_n\Omega_*\hat{B}_n - A\Omega_* B|_1 - |\hat{A}_n\Delta\hat{B}_n - A\Delta B|_1. \tag{A.19}$$

To bound (A.19), we need to bound functions of the form $|AMB - \hat{A}_n M\hat{B}_n|_1$ for arbitrary symmetric matrices $M$:

$$\begin{aligned} |AMB - \hat{A}_n M\hat{B}_n|_1 & \\ & = |(A - \hat{A}_n)MB + AM(B - \hat{B}_n) + (A - \hat{A}_n)M(\hat{B}_n - B)|_1 \\ & \leq |(A - \hat{A}_n)MB|_1 + |AM(B - \hat{B}_n)|_1 + |(A - \hat{A}_n)M(\hat{B}_n - B)|_1. && \text{(A.20)} \\ & = |(A - \hat{A}_n)A^+AMB|_1 + |AMBB^+(B - \hat{B}_n)|_1 \\ & \qquad + |(A - \hat{A}_n)A^+AMBB^+(\hat{B}_n - B)|_1, && \text{(A.21)} \\ & \leq |AMB|_1 \Big\{ |(A - \hat{A}_n)A^+|_1 + |B^+(B - \hat{B}_n)|_1 \\ & \qquad + |(A - \hat{A}_n)A^+|_1 |B^+(B - \hat{B}_n)|_1 \Big\} && \text{(A.22)} \\ & \leq |AMB|_1 \left( C_a a_n + C_b b_n + C_a C_b a_n b_n \right), && \text{(A.23)} \end{aligned}$$

where (A.20) follows from the triangle inequality; (A.21) follows from Assumption 2 and the definition of $A^+$ and $B^+$; (A.22) follows from the sub-multiplicative property of the $|\cdot|_1$ norm; and (A.23) holds with probability at least $\min\{1 - f(C_a), 1 - g(C_b)\}$. Applying (A.23) to both terms in (A.19) gives

$$V_3 \geq -2|\hat{A}_n\Omega_*\hat{B}_n - A\Omega_* B|_1 - |\hat{A}_n\Delta\hat{B}_n - A\Delta B|_1 \geq -d_n \left( 2|[A\Omega_* B]_{\mathcal{G}}|_1 + |A\Delta B|_1 \right),$$

with probability at least $\min\{1 - f(C_a), 1 - g(C_b)\}$. Plugging this bound into (A.18) gives the result.

**Lemma 5**

*Suppose Assumptions 1–3 are true, $d_n = o(1)$, the bound in Lemma 4 holds, $\lambda \leq \epsilon\kappa(Q_\epsilon\tau)^{-1}$ for $\tau > 8$, where*

$$Q_\epsilon = \left\{ \left( \frac{3}{2} + d_n \right) \xi(p, \mathcal{G}) - 2d_n \frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\epsilon} \right\}.$$

*Then for all positive and sufficiently small $\epsilon$, $\|B^+(S - \Omega_*^{-1})A^+\|_\infty \leq \lambda/2$, implies $\|\hat{\Omega} - \Omega_*\|_F \leq \epsilon$.*

**Proof of Lemma 5.** Let $\tilde{f}$ be the objective function from (4.11). Define $\tilde{D}(\Delta) = \tilde{f}(\Omega_* + \Delta) - \tilde{f}(\Omega_*)$ so that

$$\tilde{D}(\Delta) = \operatorname{tr}(S\Delta) + \log\det(\Omega_*) - \log\det(\Omega_* + \Delta) + \lambda_1 \left\{ |\hat{A}_n(\Omega_* + \Delta)\hat{B}_n|_1 - |\hat{A}_n\Omega_*\hat{B}_n|_1 \right\}.$$

As in the proof of Lemma 1, we want to show that for $\Delta \in \mathcal{B}_\epsilon$, $\inf\{\tilde{D}(\Delta) : \|\Delta\|_F \leq \epsilon\} > 0$. Applying Lemma 4 to bound $|\hat{A}_n(\Omega_*+\Delta)\hat{B}_n|_1 - |\hat{A}_n\Omega_*\hat{B}_n|_1$ and applying the same arguments as in the proof of Lemma 1 to obtain (A.12),

$$\begin{aligned}
\tilde{D}(\Delta) \geq & \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{\lambda}{2}\left( |\,[A\Delta B]_{\mathcal{G}}\,|_1 + |\,[A\Delta B]_{\mathcal{G}^c}\,|_1 \right) + \lambda\left( |\,[A\Delta B]_{\mathcal{G}^c}\,|_1 - |\,[A\Delta B]_{\mathcal{G}}\,|_1 \right) \\
& - \lambda d_n \left( |A\Delta B|_1 + 2|[A\Omega_* B]_{\mathcal{G}}|_1 \right) \\
= & \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{3}{2}\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1 + \frac{1}{2}\lambda|[A\Delta B]_{\mathcal{G}^c}|_1 - \lambda d_n \left( |A\Delta B|_1 + 2|[A\Omega_* B]_{\mathcal{G}}|_1 \right) \\
= & \frac{1}{8}\kappa\|\Delta\|_F^2 - \frac{3}{2}\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1 + \frac{1}{2}\lambda|[A\Delta B]_{\mathcal{G}^c}|_1 \\
& - \lambda d_n \left( |[A\Delta B]_{\mathcal{G}}|_1 + |[A\Delta B]_{\mathcal{G}^c}|_1 + 2|[A\Omega_* B]_{\mathcal{G}}|_1 \right) \\
= & \frac{1}{8}\kappa\|\Delta\|_F^2 - \left( \frac{3}{2} + d_n \right)\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1 \\
& + \left( \frac{1}{2} - d_n \right)\lambda|[A\Delta B]_{\mathcal{G}^c}|_1 - 2\lambda d_n|[A\Omega_* B]_{\mathcal{G}}|_1
\end{aligned} \tag{A.24}$$

and because $d_n = o(1)$ by assumption, for sufficiently large $n$, (A.24) implies

$$\begin{aligned}
\tilde{D}(\Delta) \geq & \frac{1}{8}\kappa\|\Delta\|_F^2 - \left( \frac{3}{2} + d_n \right)\lambda|\,[A\Delta B]_{\mathcal{G}}\,|_1 - 2\lambda d_n|[A\Omega_* B]_{\mathcal{G}}|_1 \\
= & \|\Delta\|_F^2 \left\{ \frac{1}{8}\kappa - \left( \frac{3}{2} + d_n \right)\frac{\lambda}{\|\Delta\|_F}\xi(p, \mathcal{G}) - 2\lambda d_n \frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\|\Delta\|_F^2} \right\} \\
= & \|\Delta\|_F^2 \left[ \frac{1}{8}\kappa - \frac{\lambda}{\|\Delta\|_F} \left\{ \left( \frac{3}{2} + d_n \right)\xi(p, \mathcal{G}) - 2d_n \frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\|\Delta\|_F} \right\} \right].
\end{aligned} \tag{A.25}$$

Since $\|\Delta\|_F = \epsilon$ and $\lambda \leq \epsilon\kappa(\tau Q_\epsilon)^{-1}$ for $\tau > 8$, where

$$Q_\epsilon = \left\{ \left( \frac{3}{2} + d_n \right) \xi(p, \mathcal{G}) - 2d_n \frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\epsilon} \right\},$$

the inequality from (A.25) implies

$$D(\Delta) \geq \epsilon^2 \left[ \frac{1}{8}\kappa - \frac{\lambda}{\epsilon} \left\{ \left( \frac{3}{2} + d_n \right) \xi(p, \mathcal{G}) - 2d_n \frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\epsilon} \right\} \right] \geq \epsilon^2 \left( \frac{1}{8}\kappa - \frac{1}{\tau}\kappa \right) > 0,$$

which establishes the desired result.

**Lemma 6**

*If the conditions of Lemma 5 are true, then $\hat{\Delta} = \tilde{\Omega} - \Omega_*$ belongs to the set*

$$\left\{ \Delta \in \mathbb{S}^{p_n} : |[A\Delta B]_{\mathcal{G}^c}|_1 \leq \frac{(3 + 2d_n)|[A\Delta B]_{\mathcal{G}}|_1 + 4d_n|[A\Omega_* B]_{\mathcal{G}}|_1}{1 - 2d_n} \right\}.$$

**Proof of Lemma 6.** Using the same arguments as in the proof of Lemma 1 from Negahban et al. (2012), and from (A.24), we have

$$\begin{aligned} 0 \leq D(\hat{\Delta}) \leq & -\frac{\lambda}{2}(1 + 2d_n) \left( |[A\Delta B]_{\mathcal{G}}|_1 + |[A\Delta B]_{\mathcal{G}^c}|_1 \right) + \lambda \left( |[A\Delta B]_{\mathcal{G}^c}|_1 - |[A\Delta B]_{\mathcal{G}}|_1 \right) \\ & - 2\lambda d_n |[A\Omega_* B]_{\mathcal{G}}|_1 \\ = & -\frac{\lambda}{2} \left\{ (3 + 2d_n)|[A\Delta B]_{\mathcal{G}}|_1 - (1 - 2d_n)|[A\Delta B]_{\mathcal{G}^c}|_1 + 4d_n|[A\Omega_* B]_{\mathcal{G}}|_1 \right\} \end{aligned}$$

so that

$$|[A\Delta B]_{\mathcal{G}^c}|_1 \leq \frac{(3 + 2d_n)|[A\Delta B]_{\mathcal{G}}|_1 + 4d_n|[A\Omega_* B]_{\mathcal{G}}|_1}{1 - 2d_n},$$

which is the desired inequality.

**Proof of Theorem 2.** Set $\lambda = K_2 \tau_2^{-1}(n^{-1} \log p)^{1/2}$ and $\epsilon = \lambda Q_\epsilon \tau_2 \kappa^{-1}$. We can simplify the expression for $\epsilon$ by solving

$$\epsilon\kappa(K_2^2 n^{-1} \log p)^{-1/2} = \left\{ \left( \frac{3}{2} + d_n \right) \xi(p, \mathcal{G}) - 2d_n \frac{|[A\Omega_* B]_{\mathcal{G}}|_1}{\epsilon} \right\},$$

or equivalently,

$$\epsilon^2\kappa(K_2^2 n^{-1} \log p)^{-1/2} - \epsilon \left\{ \left( \frac{3}{2} + d_n \right) \xi(p, \mathcal{G}) \right\} - 2d_n |[A\Omega_* B]_{\mathcal{G}}|_1 = 0. \tag{A.26}$$

Using the quadratic formula to solve (A.26) for $\epsilon$,

$$\begin{aligned}\epsilon =& \frac{K_2}{2\kappa}\sqrt{\frac{\log p}{n}}\left[\left(\frac{3}{2}+d_n\right)\xi(p,\mathcal{G})\right.\\ &\left.+\left\{\left(\frac{3}{2}+d_n\right)^2\xi^2(p,\mathcal{G})+\frac{16 d_n\kappa}{K_2}\sqrt{\frac{n}{\log p}}|[A\Omega_* B]_{\mathcal{G}}|_1\right\}^{1/2}\right]. \qquad \text{(A.27)}\end{aligned}$$

To simplify the result, we find an $\tilde{\epsilon}$ such that $\epsilon \leq \tilde{\epsilon}$. Then $\|\tilde{\Omega}-\Omega_*\|_F \leq \epsilon$ implies $\|\tilde{\Omega}-\Omega_*\|_F \leq \tilde{\epsilon}$, so $\|B^+(S-\Omega_*)A^+\|_\infty \leq \lambda/2$ also implies $\|\tilde{\Omega}-\Omega_*\|_F \leq \tilde{\epsilon}$. Viewing the square root in (A.27) as the Euclidean norm of the sum of the square root of its two terms, we use the triangle inequality to obtain

$$\epsilon \leq \frac{K_2}{\kappa}\sqrt{\frac{\log p}{n}}\left\{\left(\frac{3}{2}+d_n\right)\xi(p,\mathcal{G})+2\left(\frac{d_n\kappa}{K_2}\sqrt{\frac{n}{\log p}}|[A\Omega_* B]_{\mathcal{G}}|_1\right)^{1/2}\right\} = \tilde{\epsilon}.$$

Then, applying Lemma 5 and Lemma A.3 from Bickel and Levina (2008), there exists constants $C_3$ and $C_4$ such that for sufficiently large $n$,

$$\begin{aligned}P\left(\|\hat{\Omega}-\Omega_*\|_F \leq \tilde{\epsilon}\right) &\geq P\left(\|B^+SA^+ - B^+\Omega_*^{-1}A^+\|_\infty \leq \frac{K_2}{2\tau_2}\sqrt{\frac{\log p}{n}}\right)\\ &\geq 1 - C_3 p^{2-C_4K_2/2\tau_2}\end{aligned}$$

which establishes $(i)$ because $1 - C_3p^{2-C_4K_2/2\tau_2} \to 1$ as $K_2 \to \infty$. To establish $(ii)$, we bound $|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_*B|_1$. By the triangle inequality,

$$\begin{aligned}|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_*B|_1 &= |\hat{A}_n\tilde{\Omega}\hat{B}_n - A\tilde{\Omega}B + A\tilde{\Omega}B - A\Omega_*B|_1\\ &\leq |\hat{A}_n\tilde{\Omega}\hat{B}_n - A\tilde{\Omega}B|_1 + |A\tilde{\Omega}B - A\Omega_*B|_1 \qquad \text{(A.28)}\end{aligned}$$

and by the argument used to obtain the inequality in (A.23), $|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\tilde{\Omega}B|_1 \leq d_n|A\tilde{\Omega}B|_1$. Using this bound on the first term in (A.28),

$$|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_*B|_1 \leq d_n|A\tilde{\Omega}B|_1 + |A\tilde{\Omega}B - A\Omega_*B|_1. \qquad \text{(A.29)}$$

Then, bounding the first term in (A.29), $|A\tilde{\Omega}B|_1 = |A\tilde{\Omega}B + A\Omega_*B - A\Omega_*B|_1 \leq |A\Omega_*B|_1 +$

$|A\tilde{\Omega}B - A\Omega_* B|_1$ so that from (A.29),

$$|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_* B|_1 \leq d_n|A\Omega_* B|_1 + (d_n+1)|A\tilde{\Omega}B - A\Omega_* B|_1. \tag{A.30}$$

To bound the right term in the sum on the right hand side of (A.30), we apply Lemma 6 to $\tilde{\Delta} = \tilde{\Omega} - \Omega_*$

$$\begin{aligned}
|A\tilde{\Omega}B - A\Omega_* B|_1 &\leq |[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + |[A\tilde{\Delta}B]_{\mathcal{G}^c}|_1 \\
&\leq |[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + \frac{(3+2d_n)|[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + 2d_n|[A\Omega_* B]_{\mathcal{G}}|_1}{1-2d_n} \\
&= \frac{(1-2d_n)|[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + (3+2d_n)|[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + 2d_n|[A\Omega_* B]_{\mathcal{G}}|_1}{1-2d_n} \\
&= \frac{4|[A\tilde{\Delta}B]_{\mathcal{G}}|_1 + 2d_n|[A\Omega_* B]_{\mathcal{G}}|_1}{1-2d_n}.
\end{aligned} \tag{A.31}$$

Because $d_n = o(1)$, there exists constants $C_5$ and $C_6$ such that for some sufficiently large $n$, (A.31) implies $|A\tilde{\Omega}B - A\Omega_* B|_1 \leq C_5\|\tilde{\Omega} - \Omega_*\|_F \xi(p,\mathcal{G}) + C_6 d_n|[A\Omega_* B]_{\mathcal{G}}|_1$. Combining this with (A.30),

$$\begin{aligned}
&|\hat{A}_n\tilde{\Omega}\hat{B}_n - A\Omega_* B|_1 \\
&\qquad \leq (d_n+1)\left\{C_5\|\tilde{\Omega} - \Omega_*\|_F \xi(p,\mathcal{G}) + C_6 d_n|[A\Omega_* B]_{\mathcal{G}}|_1\right\} + d_n|[A\Omega_* B]_{\mathcal{G}}|_1. \\
&\qquad = C_5(d_n+1)\|\tilde{\Omega} - \Omega_*\|_F \xi(p,\mathcal{G}) + C_6(d_n^2 + d_n + d_n C_6^{-1})|[A\Omega_* B]_{\mathcal{G}}|_1
\end{aligned}$$

so that using $d_n = o(1)$ and the result from Theorem 2 $(i)$ for $\|\tilde{\Omega} - \Omega_*\|_F$, we obtain the result.

# Appendix B

# Supplemental Material for Chapter 3

## B.1 Sparse inverse regression elliptical t-distribution simulations

For 200 independent replications, we generated $n$ independent copies of the random vector $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$ which has the $p+q$-variate elliptical $t$-distribution with $\nu$ degrees of freedom and parameters $\mu_* \in \mathbb{R}^{p+q}$ and $\Sigma_* \in \mathbb{S}_+^{p+q}$. This parameterization was used by Muirhead (2009). We set $\mu_* = 0$ and we picked the entries in $\Sigma_*$ by specifying $\Sigma_{*XX}$, $\Sigma_{*XY}$, and $\Sigma_{*YY}$ defined through the partition of $\Sigma_*$ used in Section 3.2.1. The $(i,j)$th element of $\Sigma_{*YY}$ was $\rho_Y^{|i-j|}$ and the $(i,j)$th element of $\Delta_*$ was $0.7^{|i-j|}$. We set $\Sigma_{*XY}^{\mathrm{T}} = \Sigma_{*YY}\eta_*$, and $\Sigma_{*XX} = \Delta_* + \eta_*\Sigma_{*XY}^{\mathrm{T}}$, where $\eta_*$ was generated as it was in Section 3.5.1 with entry-wise nonzero probability $s_*$.

Results when $\nu = 3$ and $\nu = 10$ are displayed in Figure B.1 and Figure B.2, respectively. When $n = 100$, $p = 60$, and $q = 60$, the relative performance of $I_1$ was similar to the normal data generating model studied in Section 3.5.1 for both values of $\nu$. When $n = 50$, $p = 200$, $q = 200$, $I_1$ performed worse than both ridge regression estimators except when $\nu = 10$ and $\rho_Y = 0.9$, where $I_1$ outperformed R and $L_2$.

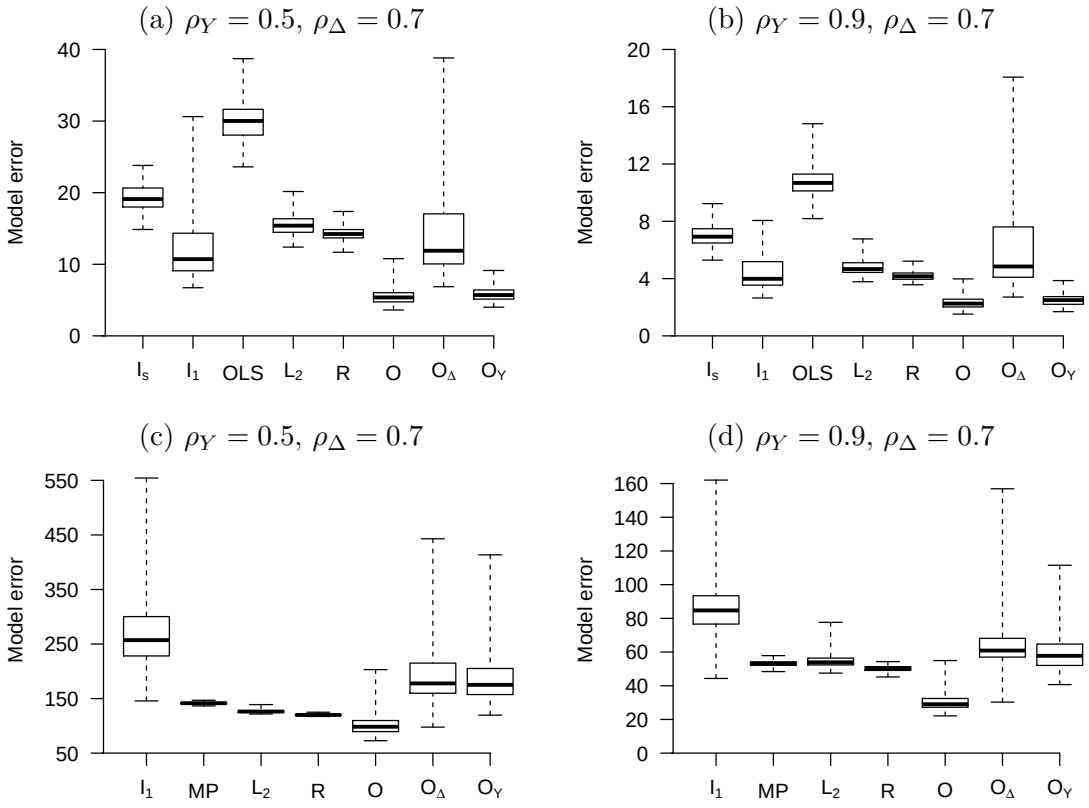



Figure B.1: Boxplots of the observed model errors from 200 replicationswhen the data generating model from Appendix B.1 is used. In (a) and (b), $n = 100, p = 60$, $q = 60$, $s_* = 0.1$, and $\nu = 3$. In (c) and (d), $n = 50, p = 200$, $q = 200$, $s_* = 0.03$, and $\nu = 3$.

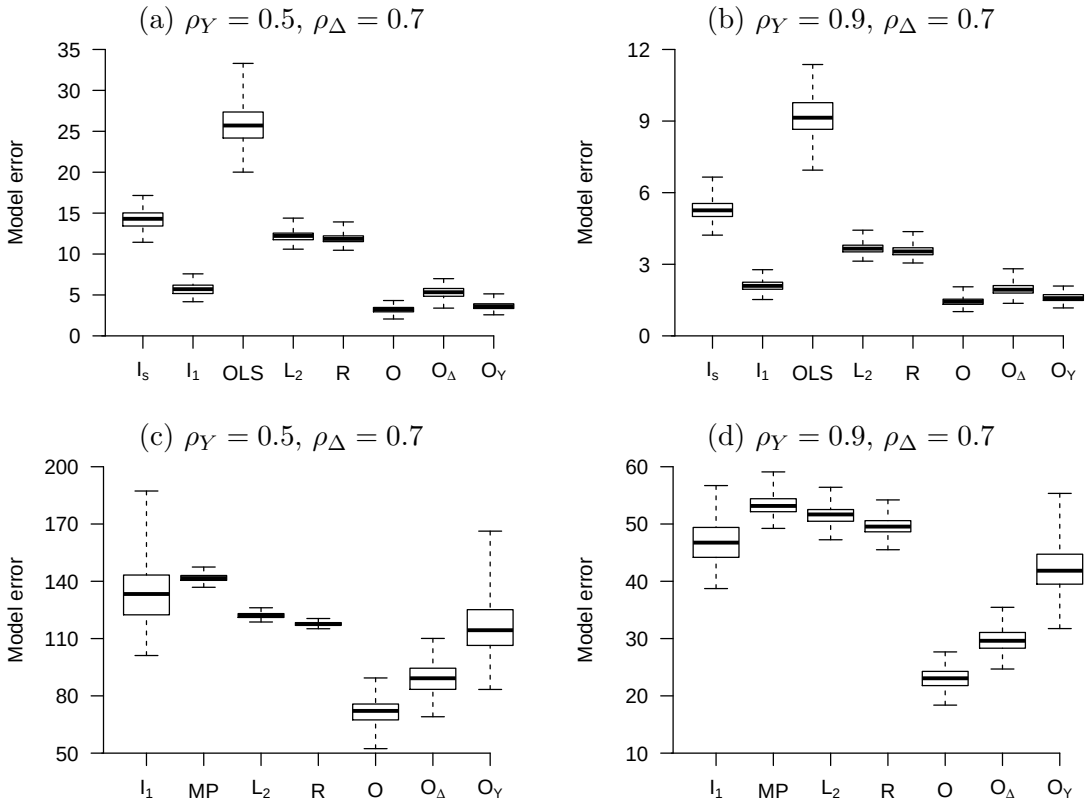



Figure B.2: Boxplots of the observed model errors from 200 replications where (a), (b) $n = 100, p = 60$, $q = 60$, $s_* = 0.1$, $\nu = 10$; (c), (d) $n = 50, p = 200$, $q = 200$, $s_* = 0.03$, $\nu = 10$; and the data generating model from Appendix B.1 is used.

## B.2 Additional sparse inverse regression simulations

In Figure B.3, we display additional side-by-side boxplots of the observed model errors for the simulation study described in Section 3.5.1. We see that $I_1$ generally outperforms the competitors. However, when $n = 50$, $p = 200$, $q = 200$, and the responses were marginally uncorrelated, there was a small number of replications in which both ridge regression estimators performed better than $I_1$ did.

## B.3 Additional Non-normal forward regression simulations

In Figure B.4, we display additional side-by-side boxplots of the observed model errors for the simulation study described in Section 3.5.2. When $n = 100$, $p = 60$, and $q = 60$, $I_1$ outperformed all non-oracle competitors. When $n = 50$, $p = 200$, $q = 200$, and the responses were marginally uncorrelated, $I_1$ outperformed the non-oracle competitors except for a small number of replications.

## B.4 Additional reduced-rank inverse regression simulation

In Figure B.5 (a)–(c), we show additional side-by-side boxplots of observed model errors for the simulation study described in Section 3.5.3. When responses are marginally uncorrelated, $I_{\text{RR}}$ outperforms the direct likelihood-based reduced-rank regression estimator, both of which performed better than the part oracle estimators $O_\Delta$ and $O_Y$.

## B.5 Additional reduced-rank forward regression simulation

In Figure B.5 (d)–(f), we display additional side-by-side boxplots of the observed model errors for the simulation study described in Section 3.5.4. In each setting, $I_{\text{RR}}$ and $I_{\text{ML}}$ performed similarly to RR. Both $I_{\text{RR}}$ and $I_{\text{ML}}$ outperformed the part oracle estimators as well. Results displayed in Figure Figure B.5 are consistent with those from Section 3.5.4. This suggests that the indirect estimators $I_{\text{ML}}$ and $I_{\text{RR}}$ are competitive with the direct likelihood-based reduced-rank regression estimator even when the inverse regression error precision matrix is not sparse.

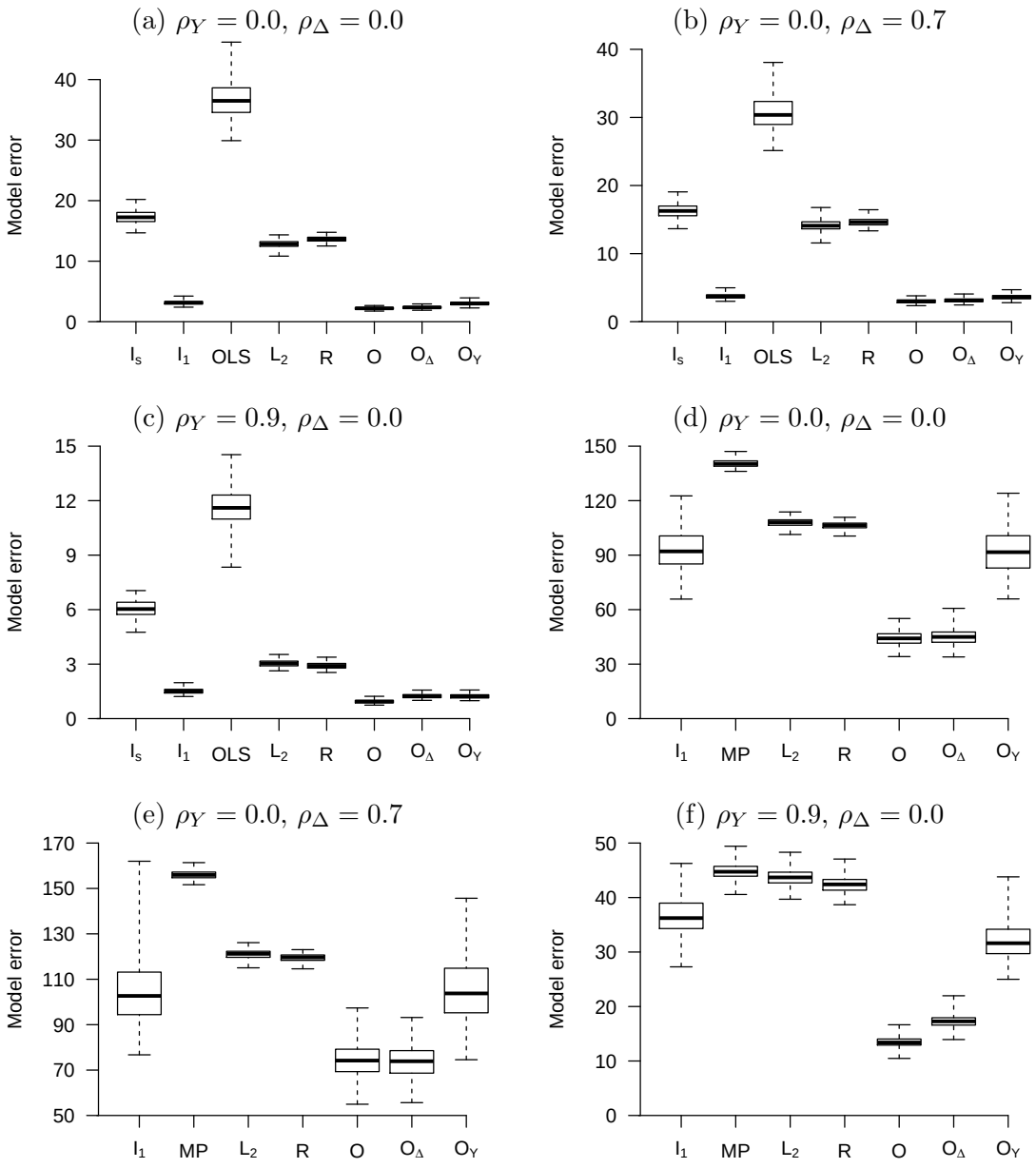



Figure B.3: Boxplots of the observed model errors from 200 replications wherethe data generating model from Section 3.5.1 was used. In (a)–(c), $n = 100, p = 60$, $q = 60$, and $s_* = 0.1$. In (d)–(f), $n = 50, p = 200$, $q = 200$, and $s_* = 0.03$.

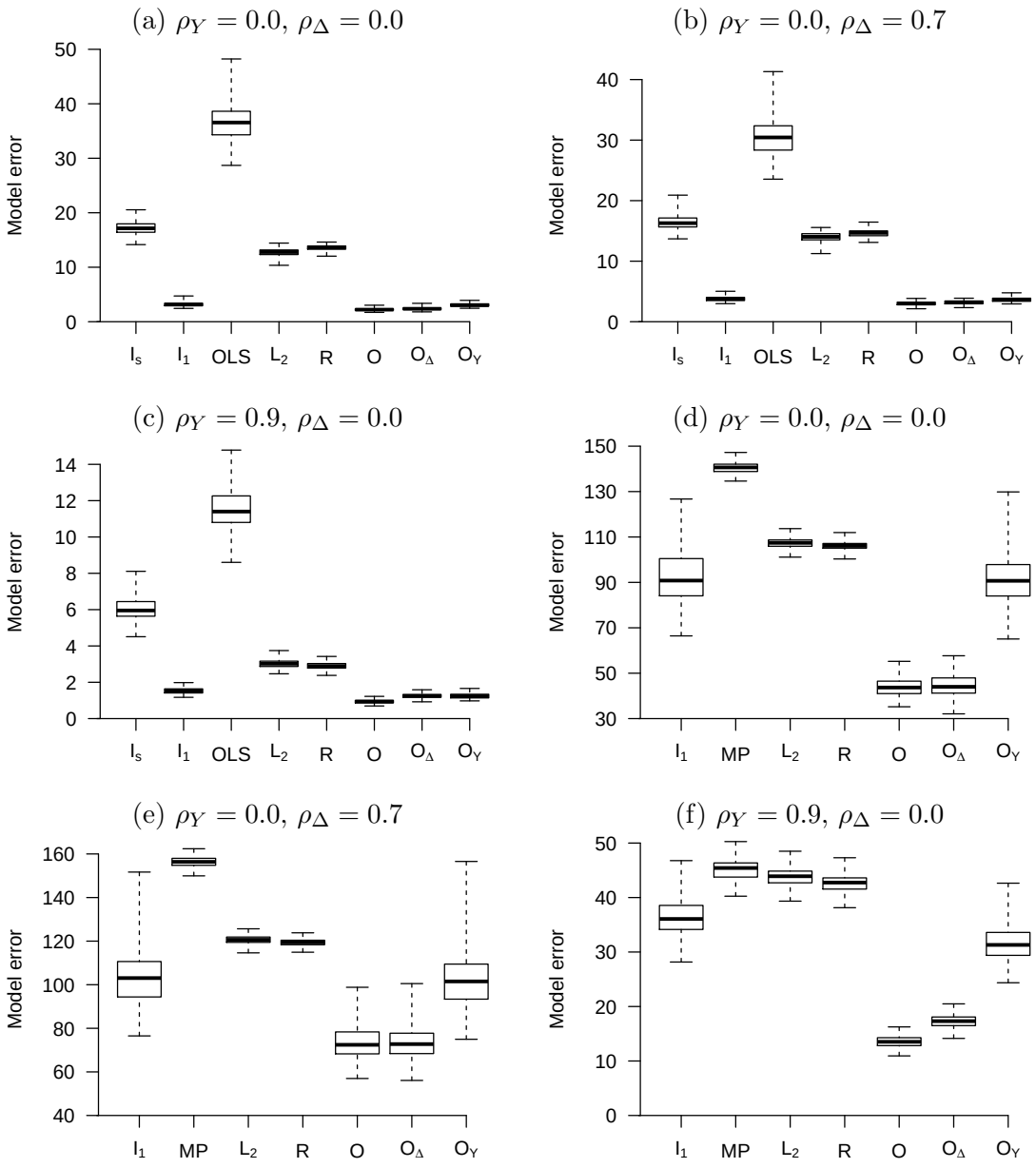



Figure B.4: Boxplots of the observed model errors from 200 replicationswhen the data generating model from Section 3.5.2 is used. In (a)–(c), $n = 100, p = 60$, $q = 60$, and $s_* = 0.1$. In (d)–(f), $n = 50, p = 200$, $q = 200$, and $s_* = 0.03$.

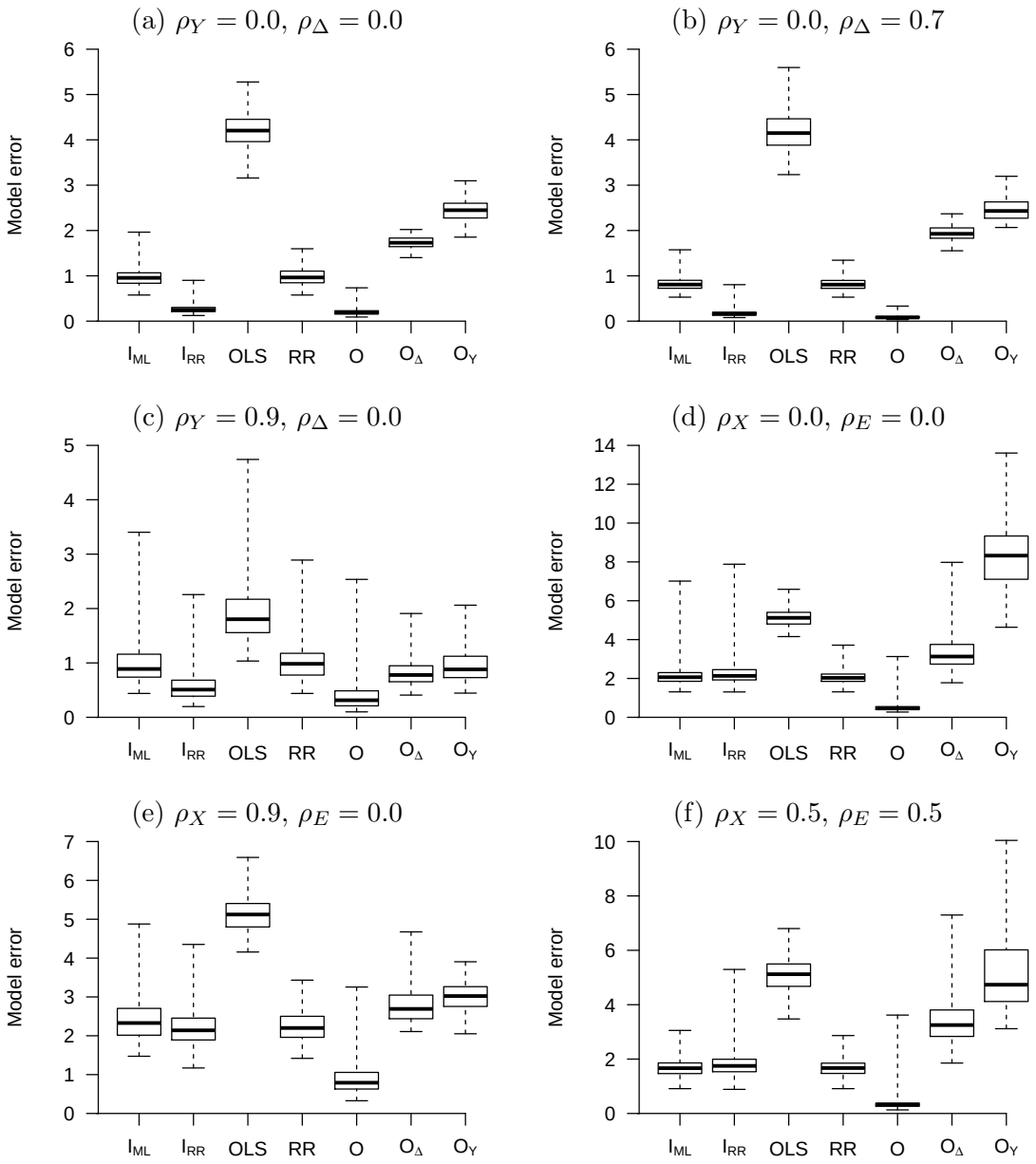



Figure B.5: Boxplots of the observed model errors from 200 replications when $n = 100, p = 20$, $q = 20$. In (a)–(d), the data generating model from Section 3.5.3 was used. In (e) and (f), the data generating model from Section 3.5.4 was used.